Three developments on spatial econometric models in data science perspective

Tohoku University Doctoral Thesis

Three developments on spatial econometric models in data science perspective

Author:
Junyue Wu

Supervisor:
Yasumasa Matsuda

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Graduate School of Economics and Management

December 23, 2020

Abstract

Recent years have seen rapid development and increasing adoption of spatial econometric models. This thesis offers three expansions tackling different subjects in this field, including a threshold extension of spatial dynamic panel data (SDPD) model, illustration of bias and inefficiency when estimating spatial models with sample data and a corresponding correction method, and a least absolute shrinkage and selection operator (LASSO) estimator for spatial weight matrix estimation.

Chapter 1 explains the motivation behind the studies included in this thesis and gives a summary of the contents in the following chapters. We introduce the spatial econometric model, then establish the three topics: a threshold extension of spatial dynamic panel data (SDPD) model, illustration of bias and inefficiency when estimating spatial models with sample data and a corresponding correction method, and a least absolute shrinkage and selection operator (LASSO) estimator for spatial weight matrix estimation.

Chapter 2 proposes a threshold extension of the spatial dynamic panel data (SDPD) model with fixed effects. We introduce a threshold variable to account for the regional dependencies of parameters in SDPD models. Moreover, we applied an extension of a unified M-estimation to estimate the parameters in the threshold SDPD models, where the consistency and asymptotic normality are established theoretically when the number of cross-sectional units tends to be infinite. The M-estimation is compared with the conditional quasi-maximum likelihood estimation by Monte Carlo experiments, showing that the M-estimation yields an estimation of less bias in cases of short time panels with robust standard errors under non-normality. We illustrate an empirical application of the threshold SDPD model to the U.S. state-level GDP and power usage growth data from 1998 to 2018, detecting the non-trivial regional dependencies of SDPD model parameters.

Chapter 3 explores bias problems when estimating spatial regression models with sample data points, and offer a solution under certain conditions. Often the individual observations used to estimate spatial regression models constitute only a sample of the theoretically observable data points. In many cases, such a sample does not even obey a specific design and it is collected only with convenience criteria as it happens e. g. when data are webscraped or crowdsourced. In this situation, we expect to observe possible biases and inefficiencies in the estimation of the spatial regression parameters. In this paper, we present the results of various Monte Carlo experiments aimed at assessing the extent of this problem in the estimation of a spatial econometric model by isolating the effects due to the sample size, those due to the pattern of the point distribution and those related to the sample criterion used in the data collection process. We also suggest an approach based on the Gibbs sampler that can be used in order to replace the unsampled data points. Our simulations and a real data case study confirm that our proposed strategy succeeds in reducing the distorting effects produced by the sample observation thus providing more reliable parameters' estimations.

Chapter 4 proposes a maximum likelihood least absolute shrinkage and selec-

tion operator (ML-LASSO) estimator for the estimation of spatial weight matrix in spatial econometric models. A cyclic coordinate descent based algorithm is used for the optimization, and the effectiveness is examined by Monte Carlo experiments. We find out that the estimator has favorable performance for arbitrary weight matrices when independent variables are present in the model, or symmetric weight matrices when there is none. An empirical use case is also illustrated with US state-level precipitation data from 1895 to 1997.

Acknowledgment

First, I would like to express my deepest appreciation to my supervisor, Professor Yasumasa Matsuda, who provides me with generous support during my graduate course. He guides me though my study with inspiring ideas and invaluable suggestions on my research. I would also like to extend my sincere thanks to Professor Giuseppe Arbia, who gave me the chance to study abroad and conduct a joint research.

Second, I am also grateful to all the people I have met in Tohoku University. I would never been able to have a memorable experience during my time here without them.

Last but not least, I would like to thank my family, whose unwavering support greatly helps me pursuing a PhD degree.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis focus on developing new methods to solve existing problems and expand the research possibilities in the field of spatial econometrics. Spatial econometrics is a relatively new sub-field of econometrics that deals with regression models with spatial interactions. Compared to previous models, the inclusion of various spatial correlations, such as spatial lag and spatial error, allows more accurate modeling thus better quality in empirical studies.

A typical spatial econometric model can be expressed as follows:

$$\begin{aligned} Y &= cl_n + X\beta + \rho WY + u, \\ u &= \alpha Wu + \varepsilon, \end{aligned} \tag{1.1}$$

where Y is an $n \times 1$ vector of observations, l_n an $n \times 1$ vector of ones, X an $n \times k$ matrix of regressors, W an $n \times n$ matrix of spatial weight matrix, u an $n \times 1$ vector of disturbance term, and ε follows independent and identically distributed (i.i.d.) error. This model is commonly refered as spatial autoregressive with additional autoregressive error stucture(SARAR, Kelejian and Prucha, 1998) or spatial autoregressive confused (SAC, LeSage and Pace, 2009) model. The spatial weight matrix W is what distinguish the model from a ordinary regression model, it represents the spatial correlation among observations Y. The spatial lag term ρWY and the spatial error term αWu are common features in the spatial models, representing the spatial stucture among observation and error respectively, where model parameter ρ and α indicate the intensity of the spatial effect.

Spatial models and their corresponding estimation methods have been developed for both cross-sectional and panel data sets in the past decades. Cliff and Ord (1973) proposed the cross-sectional spatial autoregressive (SAR) model, later also known as the spatial lag model, marking the beginning of the sub-field in econometrics. Models with different kinds of spatial interactions have been developed by Anselin (1988); Cressie (1993); Arbia (2006); LeSage and Pace (2009) since then. Meanwhile, Elhorst (2003); Baltagi et al. (2003); Kapoor et al. (2007); Lee and Yu (2010) extended spatial models to panel data with

various settings. This thesis deals with both cross-sectional and panel data in the following chapters.

Chapter 2 introduces a threshold extension of the spatial dynamic panel data (SDPD) model with fixed effects. A threshold model is a statistic model that allows its parameters to differ depending on the range of values divided by one or more threshold values in meaningful ways. Tong and Lim (1980) pioneered threshold models in time series literature by self-exciting threshold autoregressive (SETAR) models, where a lagged variable is used as a threshold on which the model switches. However, threshold models have only seen limited adaptation in spatial econometrics, Aquaro et al. (2015) applied spatial econometrics models with spatially dependent parameters, whereas Hansen (1999) extended the threshold techniques in time series to panel data, providing estimation and testing procedures for non-dynamic panels. Meanwhile, Majumdar et al. (2005) proposed a spatio-temporal model, which allows certain parameters to shift at a given time point. We believe a more generalized threshold extension and its accompanying estimation method would be a valuable addition to the spatial econometric literature.

We consider an SDPD model with a fixed effect as our base model, which can be regarded as a reduced version of the general spatial panel model summarized by Elhorst (2014). The threshold extension allows the model parameters to switch depending on predetermined groups, allowing researchers to examine regional differences. The estimation of our threshold model is based on the M-estimator proposed by Yang (2018). We have chosen this method because it incorporates a bias correction mechanism to achieve better accuracy and robustness over quasi-maximum likelihood (QML) methods in short panel settings. The M-estimator obtain the estimation result by solving the adjusted score function of the model, and since (i) the adjusted score function for the threshold model can be expressed similarly as the original linear model; (ii) the matrices composing the adjusted score in the quadratic or linear forms are still uniformly bounded in both row/column sums, we are able to expand the method to suit our model under the same set of assumptions, and prove the original theorems still stand. Also, we adapt the outer-product-of-martingale-difference (OPMD) method for the estimation of the asymptotic variance matrix accompanying the M-estimator.

Compared to past attempts, our model is based on a more generic SDPD model, along with its submodels, it offers more flexibility. Applying the threshold on the spatial terms is a first in the literature, allowing differences in the intensity of the spatial effect among regions. Moreover, the innovative M-estimator adoption makes the estimation more robust, especially for short panels.

We conducted a series of simulations to validate the proposed estimator, including comparisons between M-estimator and QML method, Gaussian and non-Gaussian errors, fixed and randomized parameters, shorter and longer panels, etc. The results show smaller bias for the M-estimator, robustness under non-Gaussian error, overall good performance under different true values of parameters, and they are comparable to Yang (2018)'s original study. Moreover, we conducted an empirical illustration using U.S. state-level GDP and power usage growth data is also conducted to demonstrate how threshold SDPD models

work to account for spatial dependencies of parameters, leading to a deeper analysis than that using usual SDPD models. We are able to identify the difference in model parameters between groups of states with different levels of income.

Chapter 3 examines the bias and inefficiency in spatial econometric model estimation when using data sets sampled from a population, and offers a Gibbs sampling based correction method to mitigate the issue. In empirical statistic analysis, estimating a model with a sample from the population is common practice, and spatial econometrics is no exception. However, apart from the usual problem that the sample may not represent the population very well, there are additional challenges in spatial analysis. In point pattern analysis literature, it is common to distinguish the situation when all points are available (known as a mapped pattern) from the situation when only a sample of them can be observed (referred to as a sample pattern. See Diggle (1983); Illian et al. (2008)). Nevertheless, the problem is usually overlooked in the field of spatial econometrics.

Some previous research has considered the missing data problem in spatial analysis. Bennett et al. (1984); Haining et al. (1984); Griffith et al. (1989) analyzed the effects of missing spatial data and compared the performances of different methods under such circumstance. More recently Arbia et al. (2016) have reappraised the problem extending the study to the effects on the estimation of a spatial regression model. They show that the presence of missing data reduces the precision of the estimates of all the regression parameters with a reduction of the efficiency which is emphasized by the presence of strong spatial correlation and by the presence of missing points that are clustered in space. The problem of missing data is well known in the statistical literature (Little, 1988; Little and Rubin, 2019; Rubin, 1976) where solutions have been suggested to replace the observations that are missing following different interpolation strategies (Dempster et al., 1977; Rubin, 2004), although with no explicit reference to the spatial data peculiarities.

The issues of using sampled spatial data have their unique characteristics. Firstly, for the missing data problem we usually assume most observations are available, but in our case it is the opposite: we only have a small portion of the population available. Secondly, unlike usual missing data problems, we need to consider the spatial structure of the observation. To address these issues, we start by examining what kind of problem it may cause.

To examine the possible consequences of observing a sample of data when the observations are distributed in space following a certain point pattern, we conducted a series of Monte Carlo simulations. We assess different combinations of spatial patterns (Complete Spatial Randomness(CSR), clustered, inhibition), sampling methods (random, quadrant, contagion, threshold), and sample proportions, and come to the conclusion that (i) biases and inefficiencies in parameters are observed in all simulations; (ii) the differences in bias are similar among different point patterns; (iii) quadrant and contagion sampling scheme has less bias and inefficiency, the threshold has the most.

Following the results, we propose a Gibbs sampling-based method to mitigate

the inaccuracies. We consider the situation that only a sample of observations is available, also the position of all data points and all independent variables are available. Under the SL model, the Gibbs sampler builds up the unknown observations with posterior samples every iteration. We validate the method with a set of Monte Carlo simulations, the results indicate that the method is effective under all four aforementioned sampling schemes. An empirical illustration is also put forward using western Tokyo land price data in 2019, showing the proposed method is able to make estimation result from sample data closer to the population one.

Chapter 4 proposed a maximum-likelihood-based least absolute shrinkage and selection operator (ML-LASSO) to estimate the spatial weight matrix in spatial lag (SL) models. In spatial analysis, the spatial weight matrix is usually pre-determined. Its choice remains a problem for the researchers, contiguity based matrices, geographic or economic distance-based matrices are some popular candidates. some previous studies have explored methods to estimate spatial weight matrices, Bhattacharjee and Jensen-Butler (2013) proposed a method to infer weight matrix in spatial error models under the assumption that the matrix is symmetric, and Beenstock and Felsenstein (2012) use moments from the sample covariance matrix to estimate the weight matrix in spatial lag models under restrictions.

LASSO is a regression method used for variable selection and regularization, and it is proposed by Tibshirani (1996) in the field of statistics. Recently, the method is gaining popularity thanks to the widespread adoption in machine learning. Under the assumption that the spatial weight matrix is sparse, which is reasonable since contiguity based or certain kinds of geographically based matrices are sparse in nature, we can use LASSO for its estimation. Ahrens and Bhattacharjee (2015) propose a two-step LASSO estimator for spatial lag models with potentially large n and reasonable T. Lam and Souza (2019) introduced a LASSO based method which allows a combination of predetermined weight matrix as well as an estimated one.

We focus on the SL model in this study and propose a cyclic coordinate descent type method to obtain the maximum likelihood LASSO (ML-LASSO) estimation result. There are two points that make this strategy more practical: (i) the partial derivative of the likelihood function with respect to an element of the spatial weight matrix is quadratic; (ii) the active set strategy can be applied. The LASSO penalty level is chosen by minimizing extended Bayesian information criteria(EBIC) proposed by Chen and Chen (2008), and the covariance matrix adaptation evolution strategy (CMA-ES, Hansen and Ostermeier (1996)) is employed to find the optimal value.

The performance of the ML-LASSO is tested by a series of Monte Carlo experiments. The specifications consist of combinations of a weight matrix with non-zero elements on super-diagonal or both super-diagonal and sub-diagonal, with spatial unit numbers of 30,50,100, sample sizes of 50, 100, 150, and with or without regressors. The results show that (i) the estimator works well overall except the case when there is no regressor and the underlying weight matrix is non-symmetric (with super-diagonal non-zero elements); (ii)the estimator

performs better with regressors; (iii) the performance improves as the sparsity and sample size increase. we also have an empirical illustration using US precipitation data from 1895 through 1997. The proposed estimator is able to identify a first-order-contiguity-like spatial pattern.

Bibliography

Ahrens, A. and Bhattacharjee, A. (2015). Two-step lasso estimation of the spatial weights matrix. *Econometrics*, 3(1):128–155.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, volume 4. Springer Science & Business Media.

Aquaro, M., Bailey, N., and Pesaran, M. H. (2015). Quasi maximum likelihood estimation of spatial models with heterogeneous coefficients. *USC-INET Research Paper*, (15-17).

Arbia, G. (2006). *Spatial econometrics: statistical foundations and applications to regional convergence*. Springer Science & Business Media.

Arbia, G., Espa, G., and Giuliani, D. (2016). Dirty spatial econometrics. *The Annals of Regional Science*, 56(1):177–189.

Baltagi, B. H., Song, S. H., and Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of econometrics*, 117(1):123–150.

Beenstock, M. and Felsenstein, D. (2012). Nonparametric estimation of the spatial connectivity matrix using spatial panel data. *Geographical Analysis*, 44(4):386–397.

Bennett, R. J., Haining, R. P., and Griffith, D. A. (1984). The problem of missing data on spatial surfaces. *Annals of the Association of American Geographers*, 74(1):138–156.

Bhattacharjee, A. and Jensen-Butler, C. (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics*, 43(4):617–634.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Cliff, A. and Ord, J. (1973). *Spatial autocorrelation*. Pion Ltd., London.

Cressie, N. A. (1993). Statistics for spatial data. Technical report, John Wiley & Sons, Inc.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. Academic Press.

Elhorst, J. P. (2003). Specification and estimation of spatial panel data models. *International regional science review*, 26(3):244–268.

Elhorst, J. P. (2014). *Spatial econometrics: from cross-sectional data to spatial panels*, volume 479. Springer.

Griffith, D. A., Bennett, R. J., and Haining, R. P. (1989). Statistical analysis of spatial data in the presence of missing observations: a methodological guide and an application to urban census data. *Environment and Planning A*, 21(11):1511–1523.

Haining, R., Griffith, D. A., and Bennett, R. (1984). A statistical approach to the problem of missing spatial data using a first-order markov model. *The Professional Geographer*, 36(3):338–345.

Hansen, B. E. (1999). Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of econometrics*, 93(2):345–368.

Hansen, N. and Ostermeier, A. (1996). Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of IEEE international conference on evolutionary computation*, pages 312–317. IEEE.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.

Kapoor, M., Kelejian, H. H., and Prucha, I. R. (2007). Panel data models with spatially correlated error components. *Journal of econometrics*, 140(1):97–130.

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1):99–121.

Lam, C. and Souza, P. C. (2019). Estimation and selection of spatial weight matrix in a spatial lag model. *Journal of Business and Economic Statistics*.

Lee, L. and Yu, J. (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics*, 154(2):165 – 185.

LeSage, J. and Pace, R. (2009). *Introduction to Spatial Econometrics*. CRC Taylor & Francis Group.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Majumdar, A., Gelfand, A. E., and Banerjee, S. (2005). Spatio-temporal change-point modeling. *Journal of Statistical Planning and Inference*, 130(1-2):149–166.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tong, H. and Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data-with discussion. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 42(3):245–292.

Yang, Z. (2018). Unified m-estimation of fixed-effects spatial dynamic models with short panels. *Journal of econometrics*, 205(2):423–447.

Chapter 2

A threshold extension of spatial dynamic panel model with fixed effects

2.1 Introduction

A panel data model with spatial interactions has been gaining more attention since the works of Baltagi et al. (2003), Kapoor et al. (2007) and Yu et al. (2008) in the field of econometrics. Many different settings, such as static or dynamic model with spatial lag (SL) or spatial error (SE) and fixed or random effect, have been explored with their corresponding estimation methods. The spatial dynamic panel data (SDPD) model with a fixed effect is one of the most popular models in spatial panel data analysis. See, for example, studies by Lee and Yu (2010) and (Elhorst, 2014, Chapter 3)

This study focuses on a threshold extension of SDPD model with fixed effects, especially in a short panel setting. Tong and Lim (1980) pioneered threshold models in time series literature by self-exciting threshold autoregressive (SETAR) models, where a lagged variable is used as a threshold on which the model switches. Let y_t be a time series and $R_1 \cup \cdots \cup R_Q = \mathbb{R}$ be a partition of real line by mutually disjoint subsets. Using a threshold variable $y_{t-d}, d > 0$, they defined SETAR for t such that $y_{t-d} \in R_q, q = 1, \ldots, Q$, as follows:

$$y_t = \phi_0^{(q)} + \sum_{i=1}^{p} \phi_i^{(q)} y_{t-i} + \varepsilon_t,$$

where ε_t is a sequence of independent and identically distributed (i.i.d.) error terms. SETAR allows a temporal dependency of parameters by a threshold variable y_{t-d}.

Extensions of threshold models from time series to spatial data have not been conducted extensively except for a few empirical studies. For instance,

Aquaro et al. (2015) applied spatial econometrics models with spatially dependent parameters, whereas Hansen (1999) extended the threshold techniques in time series to panel data, providing estimation and testing procedures for non-dynamic panels. Meanwhile, Majumdar et al. (2005) proposed a spatio-temporal model, which allows certain parameters to shift at a given time point.

This study tries a threshold extension of SDPD model with fixed effect by allowing model parameters θ to switch from one to another, depending on a threshold variable. We introduce an exogenous and time-independent variable z_i at ith spatial region as a threshold variable. Dependent on the partition $R_q, q = 1, \ldots, Q$ in which z_i is included, dependent variable y_{ti} follows a SDPD model with parameters θ_q fixed from among $\theta_1, \ldots, \theta_Q$. In other words, parameters in SDPD models can be spatially heterogeneous depending on a threshold variable, but they are time-independent. Threshold SDPD models reduce to usual SDPD when the threshold variables are constants.

The threshold variable and the corresponding threshold value need to be determined prior to the estimation due to their exogenous nature. There are several possible criteria for the choice of the threshold, such as (i) one that provide the best forecasting performance; (ii) one that shows significant differences in the estimated parameter values among the groups and so on.

We extend the unified M-estimation originally designed by Yang (2018) for usual SDPD models to that for threshold SDPD models. Yang (2018) demonstrated the consistency and asymptotic normality of the M-estimators in cases of short temporal length when the number of spatial regions tends to be infinite. We show that the M-estimation for threshold SDPD models still holds the consistency and asymptotic normality.

Additionally, we conduct Monte Carlo experiments to compare M-estimation with conditional quasi-maximum likelihood estimation (CQMLE) to demonstrate the advantage of the M-estimation over CQMLE in finite sample sizes. The empirical illustration using U.S. state-level GDP and power usage growth data is also conducted to demonstrate how threshold SDPD models work to account for spatial dependencies of parameters, leading to a deeper analysis than that using usual SDPD models. Although existing studies (Fallahi, 2011; Mahalingam and Orman, 2018) have examined the relations between economic growth and power usage, this study is the first trial to it spatial models accounting for spatial dependencies of parameters to them.

In Section 2, we set forth our threshold SDPD models as an extension of SDPD models. The M-estimation for threshold SDPD models is introduced in relation to that for SDPD models and the asymptotic results of consistency and asymptotic normality in Section 3. Monte Carlo experiments and applications to the real example are illustrated in Sections 4 and 5, respectively. Finally, Section 7 presents our conclusion.

2.2 A threshold extension of SDPD models

Let y_{ti} be a spatial panel data at period t and regional unit i for $t = 1, \ldots, T$ and $i = 1, \ldots, n$. We consider a SDPD model with a fixed effect as our base model, which can be regarded as a reduced version of the general spatial panel model summarized by Elhorst (2014). Let $W_r, r = 1, 2, 3$ be a predetermined $n \times n$ spatial weight matrix whose (i, j)th element represents the spatial correlation between units i and j; the diagonal elements are zero and the other elements are normalized to obtain a sum of 1 for every row. Then our base model is described by

$$\begin{aligned} y_{ti} =& \mu_i + \alpha_t + x_{ti}\beta + \rho y_{t-1,i} + \lambda_1 \sum_{j=1}^{n} w_{1,ij} y_{tj} + \lambda_2 \sum_{j=1}^{n} w_{2,ij} y_{t-1,j} + u_{ti}, \\ u_{ti} =& \lambda_3 \sum_{j=1}^{n} w_{3,ij} u_{tj} + v_{ti}, i = 1, \ldots, n, t = 1, \ldots, T, \end{aligned} \tag{2.1}$$

where $\{v_{ti}\}$ is i.i.d. across i and t with mean 0 and variance σ_v^2, $\{\mu_i\}$ is individual-specific effects, $\{\alpha_t\}$ is time-specific effects, and $x_{ti} = (x_{ti1}, \ldots, x_{tip})$ denotes a vector of explanatory variable at time t that does not contain any time-invariant variable because of the model identification. Moreover, ρ, λ_1, λ_2 and λ_3 are all scalar parameters reflecting the strength of spatio-temporal dependence. Theoretical and empirical studies have extensively examined the spatial dynamic panel model with fixed effect (Lee and Yu, 2010).

Some empirical studies pointed out that regional difference in the regression parameters β and the spatio-temporal ones of $\rho, \lambda_r, r = 1, 2, 3$ is often detected and suitable modeling could significantly improve the model performance (LeSage and Chih, 2018). Here we propose a threshold extension of the aforementioned model as a base to account for the regional differences. This extension is inspired by the threshold models in time series literature (Tong and Lim, 1980) where the model switches by lagged dependent variable. Despite the popularity in time series literature, threshold models have rarely been used in the field of spatial econometrics. Hansen (1999) proposed a threshold non-dynamic panel with fixed effects, allowing parameters β for regressors x_{ti} to switch between two groups. Majumdar et al. (2005) proposed a spatio-temporal model with a mean shift at certain time points.

Let $z_i, i = 1, \ldots, n$ be an exogenous time-independent univariate threshold variable and $R_r, r = 1, 2, ..., Q$ be mutually disjoint subsets that satisfy

$$R_1 \cup R_2 \cup ... \cup R_Q = \mathbb{R}.$$

Depending on which region R_q z_i falls into, we split the model into Q regimes.

The model is described as follows, for units i such that $z_i \in R_q, q = 1, \ldots, Q$,

$$\begin{aligned} y_{ti} &= \mu_i + \alpha_t + x_{ti}\beta_q + \rho_q y_{t-1,i} + \lambda_{1q} \sum_{j=1}^{n} w_{1,ij} y_{tj} \\ &\quad + \lambda_{2q} \sum_{j=1}^{n} w_{2,ij} y_{t-1,i} + u_{ti}, \\ u_{ti} &= \lambda_{3q} \sum_{j=1}^{n} w_{3,ij} u_{tj} + v_{ti}, \end{aligned} \tag{2.2}$$

where β_q, ρ_q and $\lambda_{rq}, r = 1, 2, 3$ are the parameters in the qth regime. In comparison with the existing approaches by Hansen (1999) or Majumdar et al. (2005), our model is more flexible in the sense that it allows for three spatial effects. Here we list three examples. First, by setting all $\lambda_{qr}, r = 2, 3$ to 0, we obtain the submodel that only contains the SL. Second, by setting all $\lambda_{qr}, r = 1.2$ to 0, we leave the model with only SE. Finally, by setting all λ_{q3} to 0, we obtain the submodel with SL and space-time lag (SLTL). In the case of spatially lagged independent variables $\sum_j w_{r,ij} x_{tj}$, they can be included as part of x_{ti}.

2.3 Estimation

2.3.1 M-estimation

The estimation of our threshold model is based on the M-estimator proposed by Yang (2018). We have chosen this method because it incorporates bias correction mechanism to achieve better accuracy and robustness over quasi-maximum likelihood (QML) methods, especially when T is smaller than n. Note that the threshold is determined before the estimation rather than during the estimation process. In other words, $R_1, \ldots, R_Q$ are assumed to be known in this section, although in the later empirical section we shall choose the threshold in a trial-and-error fashion.

Following the procedure of Yang (2018), we eliminate all the time-invariant terms by first-differencing (2.2). For units i such that $z_i \in R_q, q = 1, \ldots, Q$,

$$\begin{aligned} \Delta y_{ti} &= \Delta\alpha_t + \Delta x_{ti}\beta_q + \rho_q \Delta y_{t-1,i} + \lambda_{1q} \sum_{j=1}^{n} w_{1,ij} \Delta y_{tj} \\ &\quad + \lambda_{2q} \sum_{j=1}^{n} w_{2,ij} \Delta y_{t-1,j} + \Delta u_{ti}, \\ \Delta u_{ti} &= \lambda_{3q} \sum_{j=1}^{n} w_{3,ij} \Delta u_{tj} + \Delta v_{ti}, t = 2, \ldots, T. \end{aligned} \tag{2.3}$$

Let us re-express in a matrix form for $\Delta y_t = (\Delta y_{t1}, \ldots, \Delta y_{tn})'$ by $\Delta u_t = (\Delta u_{t1}, \ldots, \Delta u_{tn})'$ and $\Delta v_t = (\Delta v_{t1}, \ldots, \Delta v_{tn})'$. For the $n \times p$ explanatory

variable matrix $x_t = (x'_{t1}, \cdots, x'_{tn})'$, let $x_t^{(q)}$ be x_t whose i rows are replaced with 0 for units i such that $z_i \notin R_q$. Similarly, for the weight matrix $W_r = (w_{r,ij}).r = 1, 2, 3$, let $W_r^{(q)}$ be W_r whose ith rows replaced with 0 for units i such that $z_i \notin R_q$. Additionally, define $W_0^{(q)}$ by the identity matrix I_n whose ith rows are replaced with 0 for $z_i \notin R_q$. Notice that

$$\begin{aligned} x_t =& x_t^{(1)} + \cdots + x_t^{(Q)}, \\ I_n =& W_0^{(1)} + \cdots + W_0^{(Q)}, \\ W_r =& W_r^{(1)} + \cdots + W_r^{(Q)}, r = 1, 2, 3. \end{aligned}$$

Then we have the matrix expression for (2.3) given by

$$\begin{aligned} \Delta y_t = \Delta\alpha_t 1_n + \sum_{q=1}^{Q} \Delta x_t^{(q)} \beta_q + \left(\sum_{q=1}^{Q} \rho_q W_0^{(q)} \right) \Delta y_{t-1} \\ + \left(\sum_{q=1}^{Q} \lambda_{1q} W_1^{(q)} \right) \Delta y_t + \left(\sum_{q=1}^{Q} \lambda_{2q} W_2^{(q)} \right) \Delta y_{t-1} + \Delta u_t, \qquad (2.4) \\ \Delta u_t = \left(\sum_{q=1}^{Q} \lambda_{3q} W_3^{(q)} \right) \Delta u_t + \Delta v_t, t = 2, \ldots, T. \end{aligned}$$

Let $\Delta Y = (\Delta y'_2, \ldots, \Delta y'_T)'$, $\Delta Y_{-1} = (\Delta y'_1, \ldots, \Delta y'_{T-1})'$, and $\Delta u = (\Delta u'_2, \ldots, \Delta u'_T)'$ and $\Delta v = (\Delta v'_2, \ldots, \Delta v'_T)'$. For d_i, the $n \times (T-1)$ matrix whose ith column is 1 and 0 otherwise. Define $\Delta X_t = (\Delta x_t^{(1)}, \ldots, \Delta x_t^{(Q)}, d_{t-1})$, $\Delta X = (\Delta X'_2, \ldots, \Delta X'_T)'$ and $\beta = (\beta'_1, \ldots, \beta'_Q, \Delta\alpha_2, \ldots, \Delta\alpha_T)'$. Let $\mathbf{W}_r^{(q)} = I_{T-1} \otimes W_r^{(q)}, r = 0, 1, 2, 3, q = 1, \ldots, Q$ and

$$\begin{aligned} B_r(\lambda_r) &= I_n - \sum_{q=1}^{Q} \lambda_{rq} W_r^{(q)}, r = 1, 3, \\ B_2(\rho, \lambda_2) &= \sum_{q=1}^{Q} \left(\rho_q W_0^{(q)} + \lambda_{2q} W_2^{(q)} \right), \end{aligned}$$

for $\lambda_r = (\lambda_{r1}, \ldots, \lambda_{rQ})'$ and $\rho = (\rho_1, \ldots, \rho_Q)'$. Then we can write the model (2.4) as

$$\begin{aligned} \Delta Y = \Delta X\beta + \left(\sum_{q=1}^{Q} \rho_q \mathbf{W}_0^{(q)} \right) \Delta Y_{-1} + \left(\sum_{q=1}^{Q} \lambda_{1q} \mathbf{W}_1^{(q)} \right) \Delta Y \\ + \left(\sum_{q=1}^{Q} \lambda_{2q} \mathbf{W}_2^{(q)} \right) \Delta Y_{-1} + \Delta u, \qquad (2.5) \\ \Delta u = \left(\sum_{q=1}^{Q} \lambda_{3q} \mathbf{W}_3^{(q)} \right) \Delta u + \Delta v. \end{aligned}$$

In the following, we distinguish the true value of a parameter from its general value by adding a subscript 0, e.g. β_0 is the true value of β. Let $\rho = (\rho_1, \ldots, \rho_Q)$ and $\lambda_r = (\lambda_{r1}, \ldots \lambda_{rQ}), r = 1, 2, 3$. It is easy to see that

$$Var(\Delta u) = \sigma_{v0}^2 \left\{ C \otimes [B_3'(\lambda_{30}) B_3(\lambda_{30})]^{-1} \right\} \equiv \sigma_{v0}^2 \Omega(\lambda_{30}),$$

where C is the $(T-1) \times (T-1)$ constant matrix

$$C = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}.$$

The quasi Gaussian loglikelihood for the parameter $\psi = (\beta, \sigma_v^2, \rho, \lambda_1, \lambda_2, \lambda_3)$ in terms of $\Delta y_2, \ldots, \Delta y_T$, as if Δy_1 is exogenous, takes the form

$$\begin{aligned} l(\psi) = & -\frac{n(T-1)}{2} \log(\sigma_v^2) - \frac{1}{2} \log |\Omega(\lambda_3)| + \log |\mathbf{B}_1(\lambda_1)| \\ & - \frac{1}{2\sigma_v^2} \Delta u(\theta)' \Omega(\lambda_3)^{-1} \Delta u(\theta), \end{aligned} \tag{2.6}$$

where $\theta = (\beta, \rho, \lambda_1, \lambda_2)$, $\Delta u(\theta) = \mathbf{B}_1(\lambda_1) \Delta Y - \mathbf{B}_2(\rho, \lambda_2) \Delta Y_{-1} - \Delta X \beta$, $\mathbf{B}_1(\lambda_1) = I_{T-1} \otimes B_1(\lambda_1)$ and $\mathbf{B}_2(\rho, \lambda_2) = I_{T-1} \otimes B_2(\rho, \lambda_2)$.

Let $S(\psi) = \frac{\partial}{\partial \psi} l(\psi)$ be the conditional quasi score function. We have

$$S(\psi) = \begin{cases} \frac{1}{\sigma_v^2} \Delta X' \Omega^{-1}(\lambda_3) \Delta u(\theta), \\ \frac{1}{2\sigma_v^2} \Delta u(\theta)' \Omega^{-1}(\lambda_3) \Delta u(\theta) - \frac{n(T-1)}{2\sigma_v^2}, \\ \frac{1}{\sigma_v^2} \Delta u(\theta)' \Omega^{-1}(\lambda_3) \mathbf{W}_0^{(q)} \Delta Y_{-1}, q = 1, \ldots, Q, \\ \frac{1}{\sigma_v^2} \Delta u(\theta)' \Omega^{-1}(\lambda_3) \mathbf{W}_1^{(q)} \Delta Y - tr(\mathbf{B}_1^{-1} \mathbf{W}_1^{(q)}), q = 1, \ldots, Q, \\ \frac{1}{\sigma_v^2} \Delta u(\theta)' \Omega^{-1}(\lambda_3) \mathbf{W}_2^{(q)} \Delta Y_{-1}, q = 1, \ldots, Q, \\ \frac{1}{2\sigma_v^2} \Delta u(\theta)' (C^{-1} \otimes A_3^{(q)}(\lambda_3)) \Delta u(\theta) - (T-1) tr(G_3^{(q)}(\lambda_3)), q = 1, \ldots, Q, \end{cases} \tag{2.7}$$

where $A_3^{(q)} = W_3'^{(q)} B_3(\lambda_3) + B_3'(\lambda_3) W_3^{(q)}$ and $G_3^{(q)} = W_3^{(q)} B_3^{-1}(\lambda_3)$.

The estimator $\hat{\psi}$ given by solving the equation $S(\psi) = 0$, is equivalent to the QML estimator that maximizes (2.6). It inconsistent unless T goes to infinity, because the necessary condition for the consistency, which is given by

$$\lim_{n \to \infty} \frac{1}{nT} S(\psi_0) \xrightarrow{p} 0,$$

is not satisfied when T is fixed. Following Yang (2018)), we shall adjust the score function to overcome this inconsistency. Particularly, we shall remove the bias of the initial conditions that does not converge to 0 when T is fixed. We need the assumption to evaluate the bias on initial conditions for y_0 in (2.2).

Assumption 1 *Under model* (2.2)*, (i) the process started m periods before data collection begins, the 0th period, and (ii) if* $m \geq 1$*,* Δy_0 *is independent of future errors* $v_t, t \geq 1$*;if* $m = 0$*,* y_0 *is independent of future errors* $v_t, t \geq 1$*.*

Under Assumption 1, using (i) the error term v_{it} in 2.2, which is independent across i and t with mean 0 and variance σ_v^2, (ii) the regressor X_t, which is exogenous, and (iii) both B_{10}^{-1} and B_{30}^{-1}, we shall evaluate $ES(\psi)$, the bias term. by reducing (2.4):

$$\Delta y_t = \mathcal{B}_0 \Delta y_{t-1} + B_{10}^{-1} \Delta X_t \beta_0 + B_{10}^{-1} B_{30}^{-1} \Delta v_t, t = 2, \ldots, T, \tag{2.8}$$

where $\mathcal{B} = \mathcal{B}(\rho, \lambda_1, \lambda_2) = B_1^{-1}(\lambda_1) B_2(\rho, \lambda_2)$. After tedious but straightforward calculations, we obtain

$$\begin{aligned} E(\Delta Y_{-1} \Delta v') &= -\sigma_{v0}^2 \mathbf{D}_{-10} \mathbf{B}_{30}^{-1}, \\ E(\Delta Y \Delta v') &= -\sigma_{v0}^2 \mathbf{D}_0 \mathbf{B}_{30}^{-1}, \end{aligned}$$

where $\mathbf{D}_{-1} = \mathbf{D}_{-1}(\rho, \lambda_1, \lambda_2)$ and $\mathbf{D} = \mathbf{D}(\rho, \lambda_1, \lambda_2)$ are given by

$$\begin{aligned} \mathbf{D} \equiv \mathbf{D}(\rho, \lambda_1, \lambda_2) &= \begin{pmatrix} \mathcal{B} - 2I_n & I_n & \cdots & 0 \\ (I_n - \mathcal{B})^2 & \mathcal{B} - 2I_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{B}^{T-3}(I_n - \mathcal{B})^2 & \mathcal{B}^{T-4}(I_n - \mathcal{B})^2 & \cdots & \mathcal{B} - 2I_n \end{pmatrix} \mathbf{B}_1^{-1}, \\ \mathbf{D}_{-1} \equiv \mathbf{D}_{-1}(\rho, \lambda_1, \lambda_2) &= \begin{pmatrix} I_n & 0 & \cdots & 0 & 0 \\ \mathcal{B} - 2I_n & I_n & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathcal{B}^{T-4}(I_n - \mathcal{B})^2 & \mathcal{B}^{T-5}(I_n - \mathcal{B})^2 & \cdots & \mathcal{B} - 2I_n & I_n \end{pmatrix} \mathbf{B}_1^{-1}. \end{aligned} \tag{2.9}$$

Let $\mathbf{C} = C \otimes I_n$. It follows that, for $q = 1, \ldots, Q$,

$$\begin{aligned} E(\Delta u' \Omega_0^{-1} \mathbf{W}_0^{(q)} \Delta Y_{-1}) &= -\sigma_{v0}^2 tr(\mathbf{C}^{-1} \mathbf{D}_{-10} \mathbf{W}_0^{(q)}), \\ E(\Delta u' \Omega_0^{-1} \mathbf{W}_1^{(q)} \Delta Y) &= -\sigma_{v0}^2 tr(\mathbf{C}^{-1} \mathbf{D}_0 \mathbf{W}_1^{(q)}), \\ E(\Delta u' \Omega_0^{-1} \mathbf{W}_2^{(q)} \Delta Y_{-1}) &= -\sigma_{v0}^2 tr(\mathbf{C}^{-1} \mathbf{D}_{-10} \mathbf{W}_2^{(q)}), \end{aligned} \tag{2.10}$$

which leads to the adjusted score function, by removing the bias term for $S(\psi)$ in (2.7). The adjusted score function is as follows:

$$S^*(\psi) = \begin{cases} \frac{1}{\sigma_v^2} \Delta X' \Omega^{-1}(\lambda_3) \Delta u(\theta), \\ \frac{1}{2\sigma_v^2} \Delta u(\theta)' \Omega^{-1}(\lambda_3) \Delta u(\theta) - \frac{n(T-1)}{2\sigma_v^2}, \\ \frac{1}{\sigma_v^2} \Delta u(\theta)' \Omega^{-1}(\lambda_3) \mathbf{W}_0^{(q)} \Delta Y_{-1} + tr(\mathbf{C}^{-1} \mathbf{D}_{-1} \mathbf{W}_0^{(q)}), q = 1, \ldots, Q, \\ \frac{1}{\sigma_v^2} \Delta u(\theta)' \Omega^{-1}(\lambda_3) \mathbf{W}_1^{(q)} \Delta Y + tr(\mathbf{C}^{-1} \mathbf{D} \mathbf{W}_1^{(q)}), q = 1, \ldots, Q, \\ \frac{1}{\sigma_v^2} \Delta u(\theta)' \Omega^{-1}(\lambda_3) \mathbf{W}_2^{(q)} \Delta Y_{-1} + tr(\mathbf{C}^{-1} \mathbf{D}_{-1} \mathbf{W}_2^{(q)}), q = 1, \ldots, Q, \\ \frac{1}{2\sigma_v^2} \Delta u(\theta)' (C^{-1} \otimes A_3^{(q)}(\lambda_3)) \Delta u(\theta) - (T-1) tr(G_3^{(q)}(\lambda_3)), q = 1, \ldots, Q. \end{cases} \tag{2.11}$$

Solving $S^*(\psi) = 0$ yields the M-estimation. To simplify the process, β, σ_v^2 can be concentrated out by solving the equation for given $\delta = (\rho, \lambda_1, \lambda_2, \lambda_3)$:

$$\begin{aligned} \hat{\beta}_M(\delta) &= (\Delta X' \Omega^{-1} \Delta X)^{-1} \Delta X' \Omega^{-1} (\mathbf{B}_1 \Delta Y - \mathbf{B}_2 \Delta Y_{-1}), \\ \hat{\sigma}_{v,M}^2(\delta) &= \frac{1}{n(T-1)} \Delta \hat{u}(\delta)' \Omega^{-1} \Delta \hat{u}(\delta), \end{aligned} \tag{2.12}$$

where $\Delta\hat{u}(\delta) = \Delta u(\hat{\beta}_M(\delta), \rho, \lambda_1, \lambda_2)$. Substituting them back into (2.11) gives the concentrated adjusted score functions, which is given by, for $q = 1, \ldots, Q$,

$$S_c^*(\delta) = \begin{cases} \frac{1}{\hat{\sigma}^2_{v,M}} \Delta\hat{u}(\delta)' \Omega^{-1}(\lambda_3) \mathbf{W}_0^{(q)} \Delta Y_{-1} + tr(\mathbf{C}^{-1}\mathbf{D}_{-1}\mathbf{W}_0^{(q)}), \\ \frac{1}{\hat{\sigma}^2_{v,M}} \Delta\hat{u}(\delta)' \Omega^{-1}(\lambda_3) \mathbf{W}_1^{(q)} \Delta Y + tr(\mathbf{C}^{-1}\mathbf{D}\mathbf{W}_1^{(q)}), \\ \frac{1}{\hat{\sigma}^2_{v,M}} \Delta\hat{u}(\delta)' \Omega^{-1}(\lambda_3) \mathbf{W}_2^{(q)} \Delta Y_{-1} + tr(\mathbf{C}^{-1}\mathbf{D}_{-1}\mathbf{W}_2^{(q)}), \\ \frac{1}{2\hat{\sigma}^2_{v,M}} \Delta\hat{u}(\delta)' (C^{-1} \otimes A_3^{(q)}(\lambda_3)) \Delta\hat{u}(\delta) - (T-1) tr(G_3^{(q)}(\lambda_3)). \end{cases} \tag{2.13}$$

Solving the equations $S_c^*(\delta) = 0$, we obtain the M-estimators $\hat{\delta}_M$, from which we have $\hat{\beta}_M = \hat{\beta}_M(\hat{\delta}_M)$ and $\hat{\sigma}^2_{v,M} = \hat{\sigma}^2_{v,M}(\hat{\delta}_M)$ by (2.12).

2.3.2 Asymptotic properties

In this section we present the asymptotic properties of the M-estimator for our threshold model. The asymptotic properties of our modified M-estimator can be proved in the same way as the original, therefore we only list the assumptions and the resulting theorems. Notice (i) the subscript 0 indicates the true value; (ii) $\rho = (\rho_1, \ldots, \rho_Q)$ and $\lambda_r = (\lambda_{r1}, \ldots \lambda_{rQ}), r = 1, 2, 3$; and (iii) $\gamma_{\min}(M)$ and $\gamma_{\max}(M)$ denote the smallest and largest eigenvalues of a real symmetric matrix M respectively. In addition to Assumption 1, we present the following assumptions in addition to Assumption 1 in the last section:

Assumption 2 *The innovations v_{ti} are i.i.d. for all t and i with $E(v_{ti}) = 0$, $Var(v_{ti}) = \sigma^2_{v0}$, and $E|v_{ti}|^{4+\epsilon} < \infty$ for some $\epsilon > 0$.*

Assumption 3 *The time-varying regressors $\{x_t, t = 0, 1, ..., T\}$ are exogenous, and their values are uniformly bounded, and $\lim_{n\to\infty} \frac{1}{nT} \Delta X' \Delta X$ exists and is non-singular.*

Assumption 4 *The space $\mathbf{\Delta}$ for the parameter $\delta = (\rho, \lambda_1, \lambda_2, \lambda_3)$ is compact, and the true parameter δ_0 lies in its interior.*

Assumption 5 *(i) For $r = 1, 2, 3,$, the elements $w_{r,ij}$ of W_r are at most of order h_n^{-1}, uniformly bounded in all i and j, and $w_{r,ii} = 0$ for all i;*

(ii) $\frac{h_n}{n} \to 0$ as $n \to \infty$;

(iii) $\{W_r, r = 1, 2, 3\}$ and $\{B_{r0}^{-1}, r = 1, 3\}$ are uniformly bounded in both row and column sums;

(iv) For $r = 1, 3$, $\{B_r^{-1}\}$ are uniformly bounded in either row or column sums, uniformly bounded in λ_r in a compact parameter space $\mathbf{\Lambda}_r$, and $0 < \underline{c}_r \leq \inf_{\lambda_r \in \mathbf{\Lambda}_r} \gamma_{\min}(B_r' B_r) \leq \sup_{\lambda_r \in \mathbf{\Lambda}_r} \gamma_{\max}(B_r' B_r) \leq \overline{c}_r < \infty$, where $\gamma_{\min}$ and $\gamma_{\max}$ denote minimum and maximum eigenvalues, respectively.

Assumption 6 *For an $n \times n$ matrix Φ uniformly bounded in either row or column sums, with elements of uniform order h_n^{-1}, and an $n \times 1$ vector ϕ with elements of uniform order $h_n^{-1/2}$,*

(i) $\frac{h_n}{n}\Delta y_1'\Phi\Delta y_1 = O_p(1)$ *and* $\frac{h_n}{n}\Delta y_1'\Phi\Delta v_2 = O_p(1)$*;*

(ii) $\frac{h_n}{n}(\Delta y_1 - E(\Delta y_1))'\phi = o_p(1)$*;*

(iii) $\frac{h_n}{n}[\Delta y_1'\Phi\Delta y_1 - E(\Delta y_1'\Phi\Delta y_1)] = o_p(1)$*;*

(iv) $\frac{h_n}{n}[\Delta y_1'\Phi\Delta v_2 - E(\Delta y_1'\Phi\Delta v_2)] = o_p(1)$.

Let us define the population counterpart of (2.13). Let $\overline{S}^*(\psi) = E[S^*(\psi)]$, and partially solve $\overline{S}^*(\psi) = 0$. For a given δ, we obtain the following,

$$\begin{aligned}\overline{\beta}_M(\delta) &= (\Delta X'\Omega^{-1}\Delta X)^{-1}\Delta X'\Omega^{-1}(\mathbf{B}_1 E\Delta Y - \mathbf{B}_2 E\Delta Y_{-1}),\\ \overline{\sigma}^2_{v,M}(\delta) &= \frac{1}{n(T-1)}E[\Delta\overline{u}(\delta)'\Omega^{-1}\Delta\overline{u}(\delta)],\end{aligned} \tag{2.14}$$

where $\Delta\overline{u}(\delta) = \Delta u(\overline{\beta}_M(\delta), \rho, \lambda_1, \lambda_2) = \mathbf{B}_1\Delta Y - \mathbf{B}_2\Delta Y_{-1} - \Delta X\overline{\beta}_M(\delta)$. Substituting them back into $\overline{S}^*(\psi)$, we obtain, for $q = 1, \ldots, Q$,

$$\overline{S}^*_c(\delta) = \begin{cases} \frac{1}{\overline{\sigma}^2_{v,M}}E[\Delta\overline{u}(\delta)'\Omega^{-1}(\lambda_3)\mathbf{W}_0^{(q)}\Delta Y_{-1}] + tr(\mathbf{C}^{-1}\mathbf{D}_{-1}\mathbf{W}_0^{(q)}),\\ \frac{1}{\overline{\sigma}^2_{v,M}}E[\Delta\overline{u}(\delta)'\Omega^{-1}(\lambda_3)\mathbf{W}_1^{(q)}\Delta Y] + tr(\mathbf{C}^{-1}\mathbf{D}\mathbf{W}_1^{(q)}),\\ \frac{1}{\overline{\sigma}^2_{v,M}}E[\Delta\overline{u}(\delta)'\Omega^{-1}(\lambda_3)\mathbf{W}_2^{(q)}\Delta Y_{-1}] + tr(\mathbf{C}^{-1}\mathbf{D}_{-1}\mathbf{W}_2^{(q)}),\\ \frac{1}{2\overline{\sigma}^2_{v,M}}E[\Delta\overline{u}(\delta)'(C^{-1}\otimes A_3^{(q)}(\lambda_3))\Delta\hat{u}(\delta)] - (T-1)tr(G_3^{(q)}(\lambda_3)).\end{cases} \tag{2.15}$$

The M-estimator $\hat{\delta}_M$ is a zero of $S^*_c(\delta)$, whereas δ_0 is a zero of $\overline{S}^*_c(\delta)$, which is easy to see through $\overline{\beta}_M(\delta_0) = \beta_0$ and $\overline{\sigma}^2_{v,M}(\delta) = \sigma^2_{v0}$. Thus $\hat{\delta}_M$ is consistent for δ_0 if $\sup_{\delta\in\Delta}||S^*_c(\delta) - \overline{S}^*_c(\delta)|| \to 0$ in probability, and the identifiability condition holds. Denote $\hat{\psi}_M = (\hat{\beta}_M', \hat{\sigma}^2_{v,M}, \hat{\rho}_M', \hat{\lambda}_M')'$.

Assumption 7 $\inf_{\delta:d(\delta,\delta_0)\geq\epsilon}||\overline{S}^*_c(\delta)|| > 0$ *for every* $\epsilon > 0$*, where* $d(\delta, \delta_0)$ *is a measure of distance between* δ_0 *and* δ*.*

We have the consistency and asymptotic normality for our M-estimator as an extension of those of Yang (2018).

Theorem 1 *Suppose Assumptions 1-7 hold. Assume further that*

(i) $\gamma_{\max}[Var(\Delta Y)]$ *and* $\gamma_{\max}[Var(\Delta Y_{-1})]$ *are bounded;*

(ii) $\inf_{\delta\in\Delta}\gamma_{\min}[Var(\mathbf{B}_1\Delta Y - \mathbf{B}_2\Delta Y_{-1})] \geq \underline{c}_y > 0$.

We then obtain $n \to \infty, \hat{\psi}_M \xrightarrow{p} \psi_0$.

Theorem 2 *Under the assumptions of Theorem 1, we have, as* $n \to \infty$*,*

$$\sqrt{n(T-1)}\left(\hat{\psi}_M - \psi_0\right) \xrightarrow{D} N[0, \lim_{n\to\infty}\Sigma^{-1}(\psi_0)\Gamma(\psi_0)\Sigma'^{-1}(\psi_0)],$$

where $\Sigma(\psi_0) = -\frac{1}{n(T-1)}E[\frac{\partial}{\partial\psi'}S^*(\psi_0)]$ *and* $\Gamma(\psi_0) = \frac{1}{n(T-1)}Var[S^*(\psi_0)]$*, both assumed to exist and* $\Sigma(\psi_0)$ *to be positive definite when* n *is sufficiently large.*

2.3.3 The robust estimation for the asymptotic variance matrix

The asymptotic variance-covariance matrix of the M-estimator $\hat{\psi}_M$ is shown in Theorem 2. Statistical inference for the M-estimation requires consistent estimators for $\Sigma(\psi_0)$ and $\Gamma(\psi_0)$. As $\Sigma(\psi_0)$ is the Hessian of $S^*(\psi)$ at ψ_0, it is estimated by the Hessian at $\hat{\psi}_M$, namely by

$$\hat{\Sigma} = -\frac{1}{n(T-1)}\frac{\partial}{\partial \psi'} S^*(\hat{\psi}_M), \tag{2.16}$$

which is consistent with the similar arguments in the proof of Theorem 2. Meanwhile, it is difficult to obtain a consistent estimator for

$$\Gamma(\psi_0) = \frac{1}{n(T-1)} Var[S^*(\psi_0)],$$

as it requires giving a model for Δy_1. Following Yang (2018), who proposed the outer-product-of-martingale-difference (OPMD) method free of the initial models, we extend his method to our threshold cases.

Let us re-express our model in (2.5) with the initial value Δy_1. Under Assumption 1, we apply the reduced form Equation (2.8) to (2.5) to obtain

$$\begin{aligned} \Delta Y &= \mathcal{R}\Delta \mathbf{y}_1 + \eta + \mathbb{S}\Delta v, \\ \Delta Y_{-1} &= \mathcal{R}_{-1}\Delta \mathbf{y}_1 + \eta_{-1} + \mathbb{S}_{-\not{1}}\Delta v, \end{aligned} \tag{2.17}$$

where

$$\begin{aligned} &\Delta \mathbf{y}_1 = (\Delta y_1', \ldots, \Delta y_1')' && \mathcal{R} = diag(\mathcal{B}_0, \mathcal{B}_0^2, \ldots, \mathcal{B}_0^{T-1}) \\ &\mathcal{R}_{-1} = diag(I_n, \mathcal{B}_0, \ldots, \mathcal{B}_0^{T-2}) && \eta = \mathbb{B}\mathbf{B}_{10}^{-1}\Delta X \beta_0 \\ &\eta_{-1} = \mathbb{B}_{-1}\mathbf{B}_{10}^{-1}\Delta X \beta^0 && \mathbb{S} = \mathbb{B}\mathbf{B}_{10}^{-1}\mathbf{B}_{30}^{-1} \\ &\mathbb{S}_{-1} = \mathbb{B}_{-1}\mathbf{B}_{10}^{-1}\mathbf{B}_{30}^{-1} \end{aligned}$$

$$\mathbb{B} = \begin{pmatrix} I_N & 0 & 0 & \cdots & 0 \\ \mathcal{B}_0 & I_N & 0 & \cdots & 0 \\ \mathcal{B}_0^2 & \mathcal{B}_0 & I_N & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{B}_0^{T-2} & \mathcal{B}_0^{T-3} & \mathcal{B}_0^{T-4} & \cdots & I_N \end{pmatrix}, \quad \mathbb{B}_{-1} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ I_N & 0 & 0 & \cdots & 0 & 0 \\ \mathcal{B}_0 & I_N & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathcal{B}_0^{T-3} & \mathcal{B}_0^{T-4} & \mathcal{B}_0^{T-5} & \cdots & I_n & 0 \end{pmatrix},$$

which leads to the re-expression of the adjusted score function at ψ_0 in (2.11) given by, for $q = 1, \ldots, Q$,

$$S^*(\psi_0) = \begin{cases} \Pi_1' \Delta v, \\ \Delta v' \Phi_1 \Delta v - \frac{n(T-1)}{2\sigma_{v0}^2}, \\ \Delta v' \Psi_1^{(q)} \Delta \mathbf{y}_1 + \Pi_2'^{(q)} \Delta v + \Delta v' \Phi_2^{(q)} \Delta_v + tr(\mathbf{C}^{-1}\mathbf{D}_{-10} W_0^{(q)}), \\ \Delta v' \Psi_2^{(q)} \Delta \mathbf{y}_1 + \Pi_3'^{(q)} \Delta v + \Delta v' \Phi_3^{(q)} \Delta_v + tr(\mathbf{C}^{-1}\mathbf{D}_0 W_1^{(q)}), \\ \Delta v' \Psi_3^{(q)} \Delta \mathbf{y}_1 + \Pi_4'^{(q)} \Delta v + \Delta v' \Phi_4^{(q)} \Delta_v + tr(\mathbf{C}^{-1}\mathbf{D}_{-10} W_2^{(q)}), \\ \Delta v' \Phi_5^{(q)} \Delta v - (T-1) tr(G_{30}^{(q)}), \end{cases} \tag{2.18}$$

where

$$\Pi_1 = \frac{1}{\sigma_{v0}^2}\mathbb{C}_b\Delta X, \qquad \mathbb{C}_b = C^{-1}\otimes B_{30},$$
$$\Pi_2^{(q)} = \frac{1}{\sigma_{v0}^2}\mathbb{C}_b\mathbf{W}_0^{(q)}\eta_{-1}, \qquad \Pi_3^{(q)} = \frac{1}{\sigma_{v0}^2}\mathbb{C}_b\mathbf{W}_1^{(q)}\eta,$$
$$\Pi_4^{(q)} = \frac{1}{\sigma_{v0}^2}\mathbb{C}_b\mathbf{W}_2^{(q)}\eta_{-1}, \qquad \Phi_1 = \frac{1}{2\sigma_{v0}^2}(C^{-1}\otimes I_n),$$
$$\Phi_2^{(q)} = \frac{1}{\sigma_{v0}^2}\mathbb{C}_b\mathbf{W}_0^{(q)}\mathbb{S}_{-1}, \qquad \Phi_3^{(q)} = \frac{1}{\sigma_{v0}^2}\mathbb{C}_b\mathbf{W}_1^{(q)}\mathbb{S},$$
$$\Phi_4^{(q)} = \frac{1}{\sigma_{v0}^2}\mathbb{C}_b\mathbf{W}_2^{(q)}\mathbb{S}_{-1}, \qquad \Phi_5^{(q)} = \frac{1}{\sigma_{v0}^2}[C^{-1}\otimes(G_{30}^{\prime(q)}+G_{30}^{(q)})],$$
$$\Psi_1^{(q)} = \frac{1}{\sigma_{v0}^2}\mathbb{C}_b\mathbf{W}_0^{(q)}\mathbb{R}_{-1}, \qquad \Psi_2^{(q)} = \frac{1}{\sigma_{v0}^2}\mathbb{C}_b\mathbf{W}_1^{(q)}\mathbb{R},$$
$$\Psi_3^{(q)} = \frac{1}{\sigma_{v0}^2}\mathbb{C}_b\mathbf{W}_2^{(q)}\mathbb{R}_{-1}.$$

Define $\Psi_{t+} = \sum_{s=2}^{T}\Psi_{ts}, t = 2,\ldots,T, \Theta = \Psi_{2+}(B_{30}B_{10})^{-1}, \Delta y_1^o = B_{30}B_{10}\Delta y_1$ and $\Delta y_{1t}^* = \Psi_{t+}\Delta y_1$. For a square matrix A, let A^u, A^l, and A^d be the upper-triangular, lower-triangular and diagonal matrix of A respectively. Then $A = A^u + A^l + A^d$. Let Π_t, Ψ_{ts}, Φ_{ts} be the sub-matrices of corresponding ones partitioned according to $t, s = 2, 3, .., T$.

Then define

$$\begin{aligned} g_{1i} &= \sum_{t=2}^{T}\Pi_{it}'\Delta v_{it}, \\ g_{2i} &= \sum_{t=2}^{T}(\Delta v_{it}\Delta\xi_{it} + \Delta v_{it}\Delta v_{it}^* - \sigma_{v0}^2 d_{it}), \\ g_{3i} &= \Delta v_{2i}\Delta\zeta_i + \Theta_{ii}(\Delta v_{2i}\Delta y_{1i}^o + \sigma_{v0}^2) + \sum_{t=3}^{T}\Delta v_{it}\Delta y_{1it}^*, \end{aligned} \tag{2.19}$$

where $\Delta\xi_t = \sum_{s=2}^{T}(\Phi_{st}'^u + \Phi_{ts}^l)\Delta v_s, \Delta v_t^* = \sum_{s=2}^{T}\Phi_{ts}^d\Delta v_s$, d_{it} is the diagonal element of Φ, and $\Delta\zeta = (\Theta^u + \Theta^l)\Delta y_1^o$. Let $\mathcal{G}_{n,i}$ be the σ-field generated by $(v_{j1},\ldots,v_{jT}, j = 1,\ldots,i), i = 1,\ldots,n, n \geq 1$ and $\mathcal{F}_{n,0}$ be the σ-field generated by $(v_0, \Delta y_0)$, and define $\mathcal{F}_{n,i} = \mathcal{F}_{n,0}\otimes\mathcal{G}_{n,i}$. Then we obtain

$$\Pi'\Delta v = \sum_{i=1}^{n} g_{1i},$$
$$\Delta v'\Phi\Delta v - E(\Delta v'\Phi\Delta v) = \sum_{i=1}^{n} g_{2i},$$
$$\Delta v'\Psi\Delta\mathbf{y}_1 - E(\Delta v'\Psi\Delta\mathbf{y}_1) = \sum_{i=1}^{n} g_{3i},$$

and $\{(g'_{1i}, g_{2i}, g_{3i})', \mathcal{F}_{n,i}\}_{i=1}^n$ form a martingale difference (M.D.) sequence.

Let $g_{1ji}^{(q)}$ be g_{1i}, replacing Π with $\Pi_j^{(q)}$ for $j = 1, 2, 3, 4$ and $q = 1, \ldots, Q$. Let $g_{2ji}^{(q)}$ and $g_{3ji}^{(q)}$ be g_{2i} and g_{3i}, which are constructed in the same way. Define

$$g_i = (g'_{11i}, g_{21i}, h_{0i}^{(1)}, \ldots, h_{0i}^{(Q)}, h_{1i}^{(1)}, \ldots, h_{1i}^{(Q)}, h_{2i}^{(1)}, \ldots, h_{2i}^{(Q)}, h_{3i}^{(1)}, \ldots, h_{3i}^{(Q)})',$$

where

$$\begin{aligned} h_{0i}^{(q)} &= g_{31i}^{(q)} + g_{12i}^{(q)} + g_{22i}^{(q)}, \\ h_{1i}^{(q)} &= g_{32i}^{(q)} + g_{13i}^{(q)} + g_{23i}^{(q)}, \\ h_{2i}^{(q)} &= g_{33i}^{(q)} + g_{14i}^{(q)} + g_{24i}^{(q)}, \\ h_{3i}^{(q)} &= g_{25i}^{(q)}. \end{aligned}$$

Then

$$S^*(\psi_0) = \sum_{i=1}^n g_i,$$

and $\{g_i, \mathcal{F}_{n,i}\}$ form a vector M.D. sequence. It follows that $\Gamma(\psi_0) = Var(S^*(\psi_0))$ is estimated consistently by

$$\hat{\Gamma} = \frac{1}{n(T-1)} \sum_{i=1}^n \hat{g}_i \hat{g}_i',$$

where $\hat{g}_i$ is obtained by replacing ψ_0 with $\hat{\psi}_M$ and Δv by its observed counterpart.

Theorem 3 *Under the assumptions in 1, we have, as $n \to \infty$,*

$$\hat{\Gamma} - \Gamma(\psi_0) = \frac{1}{n(T-1)} \sum_{i=1}^n \{\hat{g}_i \hat{g}_i' - E(g_i g_i')\} \xrightarrow{p} 0,$$

and hence $\hat{\Sigma}^{-1} \hat{\Gamma} \hat{\Sigma}'^{-1} - \Sigma^{-1}(\psi_0) \Gamma(\psi_0) \Sigma'^{-1}(\psi_0) \xrightarrow{p} 0$.

2.4 Simulation

This section conducts simulations under several settings to compare the M-estimation with CQMLE. The M-estimation is obtained by solving (2.13), whereas CQMLE is obtained by solving (2.7). The model for the simulation experiments is

$$\begin{aligned} y_{ti} = \mu_i + x_{ti}\beta_q + \rho_q y_{t-1,i} + \lambda_{1q} \sum_{j=1}^n w_{1,ij} y_{tj} \\ + \lambda_{2q} \sum_{j=1}^n w_{2,ij} y_{t-1,i} + u_{ti}, \qquad (2.20) \\ u_{ti} = \lambda_{3q} \sum_{j=1}^n w_{3,ij} u_{tj} + v_{ti}, \end{aligned}$$

where x_{ti} is independently drawn from the uniform distribution on $[-1, 1]$, and the fixed effects μ_i are generated from $\frac{1}{T}\sum_{t=1}^{T} X_t + \epsilon$, where $\epsilon \sim i.i.d.\, N(0, 0.5)$. We simulated y_{tj} for t from 1 to 100 by setting $y_0 j = 0$, and use the last T periods for the studies to guarantee Assumption 1. We simulate the threshold variables of $Z_i, i == 1, 2, \ldots, n$ from i.i.d. standard normal variables, dividing $R == R1, R2$ at the origin with by Q = = 2. We established seven cases to validate the threshold SDPD model. Case 1 is designed to examine the M-estimation in comparison with CQMLE under the following setting:

C1. $n = 50, T = 5$ standard normal errors and the identical weights $W_1 = W_2 = W_3$ of the first contiguity over 5×10 grid.

Meanwhile, Cases 2-7 are established to check (1) the bias correction of the M-estimation for short panels, (2) the effects of non-Gaussian errors, and (3) the effects when the weight matrices of $W1$, $W2$ and $W3$ are not necessarily identical, under the following variety of settings:

C2. $n = 50, T = 5$, standard normal errors and the identical first contiguity weights;

C3. $n = 50, T = 5$, Gaussian mixture errors of 90% $N(0, 1)$ and 10% $N(0, 4^2)$, then row-normalized;

C4. $n = 50, T = 5$, χ_3^2 errors and the identical first contiguity weights;

C5. $n = 50, T = 5$, standard normal errors, and the identical first contiguity weights, where the model parameters are simulated uniformly on $[-0.3, 0.3]$ for each iteration;

C6. $n = 50, T = 5$ standard normal errors and the randomized weight matrices, where for every row two random elements w_{ij} $(j = 1, 2, \ldots, n$ and $j \neq i)$ are assigned 1, others 0, then row-normalized;

C7. $n = 20, T = 20$, standard normal errors and the identical first contiguity weights over 5×4 grid.

We constructed the M-estimator and CQMLE in Case 1, whereas in Cases 2-7, we constructed the M-estimator with the robust standard error, denoted as $\hat{se}$, which is obtained in Theorem 3 by the square root of the diagonal elements of $\hat{\Sigma}^{-1}\hat{\Gamma}\hat{\Sigma}'^{-1}$, where the standard error, denoted as $\hat{se}_H$, obtained by those of $\hat{\Sigma}^{-1}$, was also evaluated for comparison. Using R Core Team (2016) and Konen and Hansen (2015), we evaluated the bias and root mean squared error (RMSE) of the M-estimator with the average $\hat{se}$ and $\hat{se}_H$ from 100 iterations. The results are reported in Tables 1-7.

Table 2.1 compares the bias and RMSE of the M-estimator and CQMLE of the full model under Gaussian error. We can see that the M-estimator performed preferably for ρ_q and β, with much smaller bias. Moreover, Tables 2.2, 2.3 and 2.4 present the performance of the M-estimator under the three diferent error distributions, $\hat{se}$ and $\hat{se}_H$. Similar to the results of Yang (2018), there are no

		M-estimator		CQMLE	
	TRUE	Bias	RMSE	Bias	RMSE
λ_{11}	0.3	-0.0126	0.0677	-0.0009	0.0833
λ_{21}	0.1	0.0133	0.0742	-0.0139	0.0643
λ_{31}	0.1	0.0170	0.1992	-0.0137	0.0804
ρ_1	0.3	-0.0026	0.0337	-0.0175	0.0351
β_1	7	-0.0042	0.2225	-0.0465	0.2413
λ_{12}	0.1	0.0036	0.0702	-0.0093	0.0798
λ_{22}	0.2	-0.0134	0.0741	-0.0107	0.0642
λ_{32}	0.3	0.0036	0.1902	0.0422	0.2061
ρ_2	0.1	0.0011	0.0620	-0.0830	0.1030
β_2	3	0.0173	0.2138	-0.0931	0.2314
σ^2	1	-0.0635	0.1338	-0.0994	0.1497

Table 2.1: The M-estimator and CQMLE in Case 1: $n = 50, T = 5$ standard normal errors and the identical weights $W1 = W2 = W3$ of the first contiguity over 5×10 grid.

	TRUE	Mean	sd	Bias	RMSE	$\hat{se}$	$\hat{se}_H$
λ_{11}	0.3	0.2874	0.0669	-0.0126	0.0677	0.0840	0.0717
λ_{21}	0.1	0.1133	0.0734	0.0133	0.0742	0.1090	0.0698
λ_{31}	0.1	0.1170	0.1995	0.0170	0.1992	0.2115	0.1988
ρ_1	0.3	0.2974	0.0338	-0.0026	0.0337	0.0328	0.0321
β_1	7	6.9958	0.2236	-0.0042	0.2225	0.2105	0.2159
λ_{12}	0.1	0.1036	0.0705	0.0036	0.0702	0.0775	0.0702
λ_{22}	0.2	0.1866	0.0732	-0.0134	0.0741	0.0892	0.0700
λ_{32}	0.3	0.3036	0.1911	0.0036	0.1902	0.2014	0.1931
ρ_2	0.1	0.1011	0.0623	0.0011	0.0620	0.0705	0.0631
β_2	3	3.0173	0.2141	0.0173	0.2138	0.2144	0.2102
σ^2	1	0.9365	0.1184	-0.0635	0.1338	0.1103	0.1108

Table 2.2: The M-estimator in Case 2: $n = 50, T = 5$, standard normal errors and the identical first contiguity weights over 5×10 grid.

	TRUE	Mean	sd	Bias	RMSE	$\hat{se}$	$\hat{se}_H$
λ_{11}	0.3	0.2758	0.0747	-0.0242	0.0782	0.0836	0.0707
λ_{21}	0.1	0.1170	0.0691	0.0170	0.0708	0.1129	0.0708
λ_{31}	0.1	0.1274	0.1997	0.0274	0.2006	0.2119	0.1986
ρ_1	0.3	0.2965	0.0331	-0.0035	0.0331	0.0336	0.0325
β_1	7	6.9958	0.2142	-0.0042	0.2132	0.2229	0.2180
λ_{12}	0.1	0.1070	0.0777	0.0070	0.0776	0.0737	0.0708
λ_{22}	0.2	0.1861	0.0617	-0.0139	0.0630	0.0884	0.0719
λ_{32}	0.3	0.3070	0.2091	0.0070	0.2082	0.2036	0.1977
ρ_2	0.1	0.1179	0.0771	0.0179	0.0788	0.0705	0.0651
β_2	3	3.0150	0.2348	0.0150	0.2341	0.1996	0.2114
σ^2	1	0.9637	0.2038	-0.0363	0.2060	0.2153	0.1143

Table 2.3: The M-estimator in Case 3: $n = 50, T = 5$, Gaussian mixture errors of 90% $N(0, 1)$ and 10% $N(0, 4^2)$, and the identical first contiguity weights over 5×10 grid.

significant differences exist between $\hat{se}$ and $\hat{se}_H$ for Gaussian errors, whereas ses of for the other two non-Gaussian error cases are much closer to the corresponding RMSEs, which shows the robustness under non-normality. Table 2.5 reports the results under randomized parameters, The biases and RMSEs are comparable with those in Table 2.2, which indicates that the M-estimator works well for different sets of parameters. Table 2.6 presents the results for randomized W_1 , W_2, and W_3, which are also comparable with those of Table 2.2, indicating that identical choice of W_1 , W_2, and W_3 does not necessarily affect the estimation performance. Finally, the results of Table 2.7 show similar biases with those in

	TRUE	Mean	sd	Bias	RMSE	$\hat{se}$	$\hat{se}_H$
λ_{11}	0.3	0.2739	0.0790	-0.0261	0.0828	0.0789	0.0707
λ_{21}	0.1	0.0863	0.0749	-0.0137	0.0758	0.1093	0.0690
λ_{31}	0.1	0.1072	0.1978	0.0072	0.1970	0.2100	0.2023
ρ_1	0.3	0.3008	0.0335	0.0008	0.0334	0.0333	0.0325
β_1	7	6.9827	0.2366	-0.0173	0.2361	0.2099	0.2161
λ_{12}	0.1	0.1153	0.0771	0.0153	0.0783	0.0761	0.0693
λ_{22}	0.2	0.2070	0.0667	0.0070	0.0667	0.0903	0.0705
λ_{32}	0.3	0.2925	0.2020	-0.0075	0.2011	0.2036	0.1954
ρ_2	0.1	0.1006	0.0666	0.0006	0.0663	0.0674	0.0621
β_2	3	2.9900	0.2340	-0.0100	0.2330	0.2106	0.2070
σ^2	1	0.9391	0.1819	-0.0609	0.1910	0.1592	0.1112

Table 2.4: The M-estimator in Case 4: $n = 50, T = 5$, χ^2_3 errors and the identical first contiguity weights over 5×10 grid.

	TRUE	Mean	sd	Bias	RMSE
λ_{11}	$-0.3 \sim 0.3$	-	-	-0.0090	0.0736
λ_{21}	$-0.3 \sim 0.3$	-	-	0.0010	0.0715
λ_{31}	$-0.3 \sim 0.3$	-	-	-0.0069	0.2208
ρ_1	$-0.3 \sim 0.3$	-	-	0.0020	0.0317
β_1	7	6.9736	0.2241	-0.0264	0.2245
λ_{12}	$-0.3 \sim 0.3$	-	-	-0.0138	0.0656
λ_{22}	$-0.3 \sim 0.3$	-	-	0.0022	0.0753
λ_{32}	$-0.3 \sim 0.3$	-	-	0.0226	0.1715
ρ_2	$-0.3 \sim 0.3$	-	-	0.0197	0.0682
β_2	3	2.9886	0.1899	-0.0114	0.1893
σ^2	1	0.9400	0.0971	-0.0600	0.1137

Table 2.5: The M-estimator in Case 5: $n = 50, T = 5$, standard normal errors, and the identical first contiguity weights over 5×10 grid, where the model parameters are simulated uniformly on $[-0.3, 0.3]$ for each iteration.

	TRUE	Mean	sd	Bias	RMSE	$\hat{se}$	$\hat{se}_H$
λ_{11}	0.3	0.2817	0.0545	-0.0183	0.0573	0.0545	0.0521
λ_{21}	0.1	0.1006	0.0490	0.0006	0.0487	0.0737	0.0516
λ_{31}	0.1	0.1214	0.1584	0.0214	0.1591	0.1713	0.1663
ρ_1	0.3	0.2982	0.0345	-0.0018	0.0343	0.0320	0.0305
β_1	7	6.9839	0.2257	-0.0161	0.2251	0.2097	0.2095
λ_{12}	0.1	0.1000	0.0516	0.0000	0.0513	0.0568	0.0523
λ_{22}	0.2	0.1840	0.0520	-0.0160	0.0542	0.0604	0.0515
λ_{32}	0.3	0.3170	0.1574	0.0170	0.1576	0.1705	0.1677
ρ_2	0.1	0.1039	0.0670	0.0039	0.0668	0.0677	0.0621
β_2	3	3.0245	0.2048	0.0245	0.2052	0.2125	0.2085
σ^2	1	0.9378	0.1051	-0.0622	0.1217	0.1087	0.1098

Table 2.6: The M-estimator in Case 6: $n = 50, T = 5$ standard normal errors and the randomized weight matrices, where for every row two random elements w_{ij} ($j = 1, 2, \ldots, n$ and $j \neq i$) are assigned 1, others 0, then row-normalized.

Table 2.2, although the RMSEs are improved because of the longer period of T. The results for submodels yield similar performance, which are shown in 2.A.4.

2.5 Empirical example

As an applied illustration, we examined state GDP and power usage growth based on panel data for 48 conterminous United States during 1998-2018. Mahalingam and Orman (2018) uses this data set[1] to examine the Granger causality between

[1]The original study uses panel data from 1978 to 2014, but the authors mention that the U.S. Bureau of Economic Analysis(BEA) warns against combining GDP data before and after

	TRUE	Mean	sd	Bias	RMSE	$\hat{se}$	$s\hat{e}_H$
λ_{11}	0.3	0.2939	0.0488	-0.0061	0.0490	0.0365	0.0406
λ_{21}	0.1	0.1011	0.0425	0.0011	0.0423	0.0438	0.0404
λ_{31}	0.1	0.0950	0.1259	-0.0050	0.1253	0.1130	0.1209
ρ_1	0.3	0.2991	0.0192	-0.0009	0.0191	0.0171	0.0182
β_1	7	7.0064	0.1464	0.0064	0.1458	0.1188	0.1321
λ_{12}	0.1	0.1049	0.0378	0.0049	0.0380	0.0376	0.0394
λ_{22}	0.2	0.1961	0.0475	-0.0039	0.0474	0.0387	0.0388
λ_{32}	0.3	0.2983	0.1265	-0.0017	0.1259	0.1076	0.1189
ρ_2	0.1	0.1001	0.0437	0.0001	0.0435	0.0351	0.0373
β_2	3	3.0002	0.1408	0.0002	0.1401	0.1243	0.1299
σ^2	1	0.9670	0.0710	-0.0330	0.0780	0.0682	0.0727

Table 2.7: The M-estimator in Case 7: $n = 20, T = 20$, standard normal errors and the identical first contiguity weights over 5×4 grid.

GDP and power consumption. They concluded that the causality exists but varies among the states, suggesting spatial dependencies of the parameters on states and possible applications of threshold SDPD models. We obtain the Real State Gross Domestic Product (GDP, in millions) data from the Bureau of Economic Analysis (BEA) and total energy consumption from all sources (POW, in billions of BTUs) from the U.S. Energy Information Administration. Moreover, we transform these data into the percentage growth with first difference of logarithm. Table 2.8 reports the basic statistics of the data set.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
GDP_g	-0.0920	0.0072	0.0208	0.0203	0.0340	0.2023
POW_g	-0.1758	-0.0153	0.0068	0.0046	0.0270	0.1296

Table 2.8: Descriptive statistics of U.S state-level GDP and power consumption growth, 1998-2018.

We fit the threshold SDPD model with Z_i, the threshold variables, given by state-level income per capita in 1998 (obtained from BEA). Dividing R into two regions, R_1 and R_2, at the median of $z_1, \ldots, z_n$, we fit the threshold SDPD model given by, for $z_i \in R_q, q = 1, 2$, the following:

$$\begin{aligned} GDP_{g,ti} &= \alpha_t + POW_{g,ti}\beta_q + \rho_q GDP_{g,t-1,i} + \lambda_{1q}\sum_{j=1}^{N} w_{ij} GDP_{g,tj} \\ &\quad + \lambda_{2q}\sum_{j=1}^{N} w_{ij} GDP_{g,t-1,i} + u_{ti}, \quad q = 1, 2, \\ u_{ti} &= \lambda_{3q}\sum_{j=1}^{N} w_{ij} u_{tj} + \varepsilon_{ti}, \quad q = 1, 2, \\ \varepsilon_{ti} &\sim N(0, \sigma^2). \end{aligned} \tag{2.21}$$

where GDP_g and POW_g are the percentage growth of GDP and POW, respec-

the year 1997. We choose to follow BEA's suggestion and only use the data after 1997.

tively, α_t is the fixed effect for time period t, w_{ij} is the elements of the spatial weight matrix defined by first-order contiguity of the states. See Figure 2.1 for the two groups of states separated by the threshold.

With Theorem 2 and 3, we can conduct Wald test to see whether there are significant differences in estimated parameters between two groups. For parameter $\hat{\tau}, \hat{\tau} = \hat{\lambda_1}, \hat{\lambda_2}, \hat{\lambda_3}, \hat{\rho}, \hat{\beta}$, we have the null hypothesis $H_0 : \hat{\tau_1} = \hat{\tau_2}$ and the alternative hypothesis $H_1 : \hat{\tau_1} \neq \hat{\tau_2}$. Under the null hypothesis, $\hat{\tau_1} - \hat{\tau_2}$ follows $N(0, Var(\hat{\tau_1}) + Var(\hat{\tau_2}) - 2Cov(\hat{\tau_1}, \hat{\tau_2}))$.

	FULL	SL	SE	SLTL
λ_1	0.1212	0.3582		0.3588
	(0.0614)	(0.0487)		(0.0494)
λ_2	0.0046			0.0141
	(0.0738)			(0.0664)
λ_3	0.2437		0.3580	
	(0.0584)		(0.0509)	
ρ	-0.0672	-0.0653	-0.0662	-0.0684
	(0.0370)	(0.0352)	(0.0370)	(0.0375)
β	0.2419	0.2420	0.2359	0.2414
	(0.0678)	(0.0654)	(0.0693)	(0.0653)
σ^2	0.0006	0.0006	0.0006	0.0006
	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Table 2.9: US state-level GDP and power consumption growth, 1998-2018, non-threshold models

	FULL	SL	SE	SLT
λ_{11}	0.1159	0.3836		0.3780
	(0.1049)	(0.0598)		(0.0636)
λ_{21}	-0.0886			-0.0731
	(0.0837)			(0.0752)
λ_{31}	0.3408		0.4322	
	(0.1365)		(0.1134)	
ρ_1	0.0483	0.0062	0.0087	0.0470
	(0.0406)	(0.0446)	(0.0442)	(0.0397)
β_1	0.1966	0.1710	0.1673	0.1803
	(0.0813)	(0.0794)	(0.0804)	(0.0766)
λ_{12}	0.1753	0.3348		0.3431
	(0.1214)	(0.0425)		(0.0657)
λ_{22}	0.0860			0.0776
	(0.0828)			(0.0710)
λ_{32}	0.0955		0.2800	
	(0.1666)		(0.0969)	
ρ_2	-0.1529	-0.1239	-0.1162	-0.1557
	(0.0470)	(0.0408)	(0.0398)	(0.0453)
β_2	0.2665	0.2879	0.2763	0.2806
	(0.0829)	(0.0770)	(0.0751)	(0.0792)
σ^2	0.0006	0.0006	0.0006	0.0006
	(0.0001)	(0.0001)	(0.0001)	(0.0001)

Table 2.10: US state-level GDP and power consumption growth, 1998-2018, threshold models

	FULL	SL	SE	SLT
λ_1	0.5732	0.4874		0.6775
λ_2	0.0490			0.0533
λ_3	0.3243		0.4156	
ρ	0.0010	0.0249	0.0153	0.0006
β	0.5001	0.2179	0.1818	0.2995

Table 2.11: US state-level GDP and power consumption growth, 1998-2018, two-sided Wald test, p-values

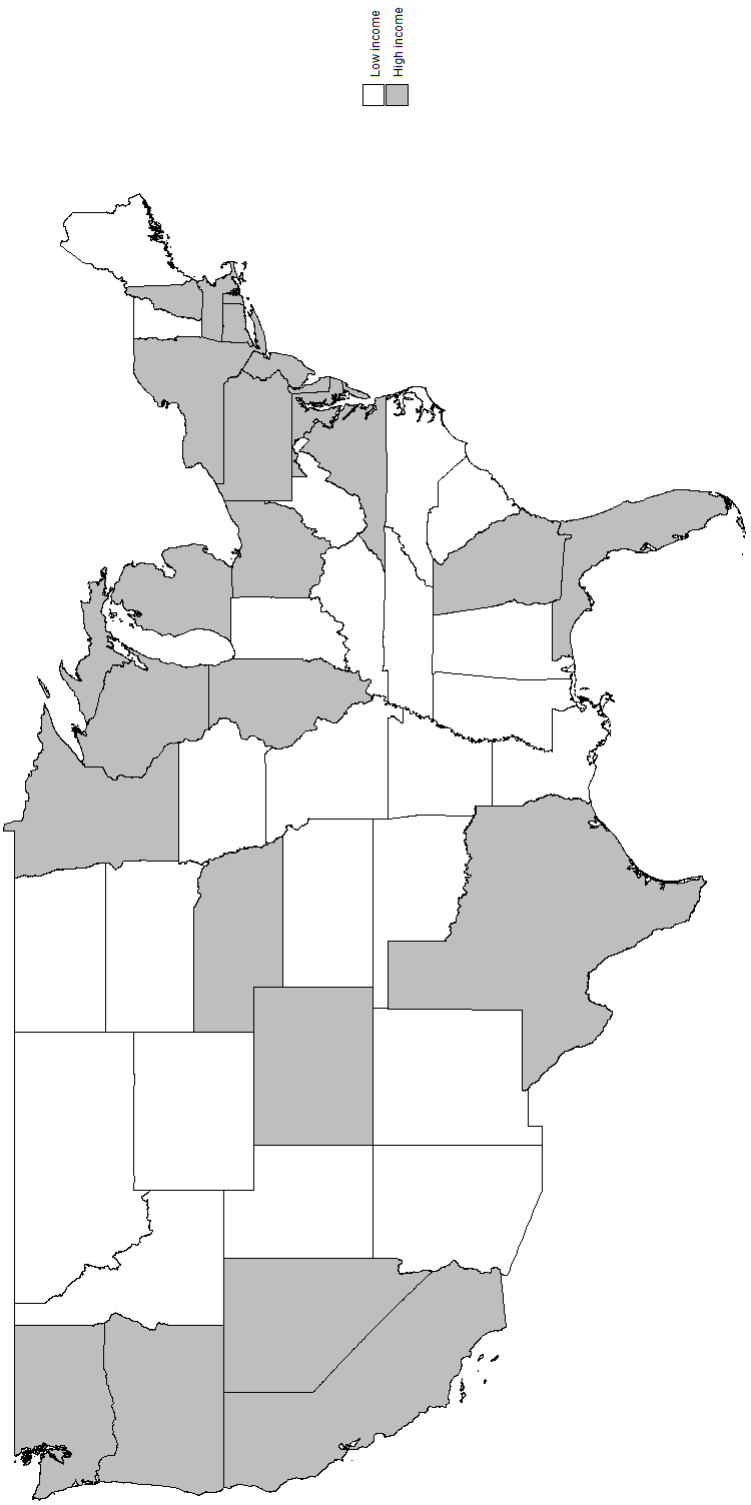

Figure 2.1: US state-level income per capita in 1998, by median

Table 2.9 and 2.10 report the estimation results of the non-threshold (2.1) and threshold SDPD models (2.2) respectively. Table 2.11 presents the p-values of the two-sided Wald test for estimated parameters in the threshold model. The non-threshold result presents the existence of spatial effect in addition to the correlation between power usage and GDP growth revealed by previous studies of Mahalingam and Orman (2018). The results of the threshold models give us further insight, showing significant difference in spatio-temporal lag and dynamic parameters between the two groups of regions, with stronger spatio-temporal correlation for regions with higher income and significant negative temporal correlation for lower income regions. Meanwhile, we can see that power consumption growth tends to have a larger impact on GDP growth in regions with lower income, although the difference is not statistically significant.

2.6 Conclusion

We introduced a threshold extension of SDPD model with fixed effects to account for spatial dependencies of parameters often observed in several empirical studies. Adapting the M-estimation for SDPD models of Yang (2018) to the threshold extension, we proposed the M-estimation to correct the bias of CQMLE in cases of short time panels. The simulation experiments reveal that the M-estimation has less bias with the standard error robust against non-normality of error terms. The empirical application to U.S GDP and power consumption successfully identifies the spatial dependencies of the parameters on per-capita income to clarify the relationship between them.

One significant restriction in the paper is that a threshold needs to be known and time-invariant to guarantee the asymptotic properties of the M-estimation. Time-varying threshold that may bring more possibilities is left for future studies.

2.A Appendix

2.A.1 Proof of Theorem 1

By the equations in (2.13) and (2.15), we have, for $q = 1, \ldots, Q$,

$$S_c^*(\delta) - \overline{S}_c^*(\delta) =$$
$$\begin{cases} \frac{1}{\hat{\sigma}^2_{v,M}} \Delta\hat{u}(\delta)' \Omega^{-1}(\lambda_3) \mathbf{W}_0^{(q)} \Delta Y_{-1} - \frac{1}{\overline{\sigma}^2_{v,M}} E[\Delta\overline{u}(\delta)' \Omega^{-1}(\lambda_3) \mathbf{W}_0^{(q)} \Delta Y_{-1}, \\ \frac{1}{\hat{\sigma}^2_{v,M}} \Delta\hat{u}(\delta)' \Omega^{-1}(\lambda_3) \mathbf{W}_1^{(q)} \Delta Y - \frac{1}{\overline{\sigma}^2_{v,M}} E[\Delta\overline{u}(\delta)' \Omega^{-1}(\lambda_3) \mathbf{W}_1^{(q)} \Delta Y], \\ \frac{1}{\hat{\sigma}^2_{v,M}} \Delta\hat{u}(\delta)' \Omega^{-1}(\lambda_3) \mathbf{W}_2^{(q)} \Delta Y_{-1} - \frac{1}{\overline{\sigma}^2_{v,M}} E[\Delta\overline{u}(\delta)' \Omega^{-1}(\lambda_3) \mathbf{W}_2^{(q)} \Delta Y_{-1}], \\ \frac{1}{2\hat{\sigma}^2_{v,M}} \Delta\hat{u}(\delta)' (C^{-1} \otimes A_3^{(q)}(\lambda_3)) \Delta\hat{u}(\delta) - \frac{1}{2\overline{\sigma}^2_{v,M}} E[\Delta\overline{u}(\delta)' (C^{-1} \otimes A_3^{(q)}(\lambda_3)) \Delta\hat{u}(\delta)]. \end{cases}$$

Under Assumption 7, the consistency of $\hat{\delta}_M$ follows from, for $q = 1, \ldots, Q$,

(a) $\inf_{\delta \in \Delta} \overline{\sigma}^2_{v,M}(\delta)$ is bounded away from 0,

(b) $\sup_{\delta\in\Delta}\left|\hat{\sigma}^2_{v,M}(\delta)-\overline{\sigma}^2_{v,M}(\delta)\right|=o_p(1)$,

(c) $\sup_{\delta\in\Delta}\frac{1}{n(T-1)}|\Delta\hat{u}(\delta)'\Omega^{-1}(\lambda_3)\mathbf{W}_0^{(q)}\Delta Y_{-1}-E[\Delta\overline{u}(\delta)'\Omega^{-1}(\lambda_3)\mathbf{W}_0^{(q)}\Delta Y_{-1}]|=o_p(1)$,

(d) $\sup_{\delta\in\Delta}\frac{1}{n(T-1)}|\Delta\hat{u}(\delta)'\Omega^{-1}(\lambda_3)\mathbf{W}_1^{(q)}\Delta Y-E[\Delta\overline{u}(\delta)'\Omega^{-1}(\lambda_3)\mathbf{W}_1^{(q)}\Delta Y]|=o_p(1)$,

(e) $\sup_{\delta\in\Delta}\frac{1}{n(T-1)}|\Delta\hat{u}(\delta)'\Omega^{-1}(\lambda_3)\mathbf{W}_2^{(q)}\Delta Y_{-1}-E[\Delta\overline{u}(\delta)'\Omega^{-1}(\lambda_3)\mathbf{W}_2^{(q)}\Delta Y_{-1}]|=o_p(1)$,

(f) $\sup_{\delta\in\Delta}\frac{1}{n(T-1)}|\Delta\hat{u}(\delta)'(C^{-1}\otimes A_3^{(q)}(\lambda_3))\Delta\hat{u}(\delta)-E[\Delta\overline{u}(\delta)'(C^{-1}\otimes A_3^{(q)}(\lambda_3))\Delta\hat{u}(\delta)]|=o_p(1)$.

With the following notations:

$$\begin{aligned}P&=\Omega^{-1/2}\Delta X(\Delta X'\Omega^{-1}\Delta X)^{-1}\Delta X'\Omega^{-1/2},\\ M&=I_n-P,\\ \tilde{\mathbf{B}}_1&=\Omega^{-1/2}\mathbf{B}_1,\tilde{\mathbf{B}}_2=\Omega^{-1/2}\mathbf{B}_2,\end{aligned}$$

we have the identities:

$$\Omega^{-1/2}\Delta\hat{u}(\delta)=M(\tilde{\mathbf{B}}_1\Delta Y-\tilde{\mathbf{B}}_2\Delta Y_{-1}), \tag{2.22}$$

$$\Omega^{-1/2}\Delta\overline{u}(\delta)=M(\tilde{\mathbf{B}}_1E\Delta Y-\tilde{\mathbf{B}}_2E\Delta Y_{-1})+\tilde{\mathbf{B}}_1(\Delta Y-E\Delta Y)-\tilde{\mathbf{B}}_2(\Delta Y_{-1}-E\Delta Y_{-1}) \tag{2.23}$$

$$=M(\tilde{\mathbf{B}}_1\Delta Y-\tilde{\mathbf{B}}_2\Delta Y_{-1})+P\left\{\tilde{\mathbf{B}}_1(\Delta Y-E\Delta Y)-\tilde{\mathbf{B}}_2(\Delta Y_{-1}-E\Delta Y_{-1})\right\}. \tag{2.24}$$

Proof of (a). From (2.23), we have

$$\begin{aligned}\overline{\sigma}^2_{v,M}=&\frac{1}{n(T-1)}tr\left[Var(\tilde{\mathbf{B}}_1\Delta Y-\tilde{\mathbf{B}}_2\Delta Y_{-1})\right]\\&+\frac{1}{n(T-1)}(\tilde{\mathbf{B}}_1E\Delta Y-\tilde{\mathbf{B}}_2E\Delta Y_{-1})'M(\tilde{\mathbf{B}}_1E\Delta Y-\tilde{\mathbf{B}}_2E\Delta Y_{-1}).\end{aligned}$$

The second term is non-negative uniformly in $\delta\in\Delta$, since M is non-negative definite. The first term is evaluated as

$$\begin{aligned}&\frac{1}{n(T-1)}tr\left[\Omega^{-1}Var(\mathbf{B}_1\Delta Y-\mathbf{B}_2\Delta Y_{-1})\right]\\ \geq&\frac{1}{n(T-1)}\gamma_{min}(C^{-1})\gamma_{min}(B_3'B_3)tr\left[Var(\mathbf{B}_1\Delta Y-\mathbf{B}_2\Delta Y_{-1}\right]>c>0,\end{aligned}$$

uniformly in $\delta\in\Delta$, by Assumption 5.

Proof of (b). By (2.22), we have

$$\hat{\sigma}^2_{v,M}(\delta)=\frac{1}{n(T-1)}(\tilde{\mathbf{B}}_1\Delta Y-\tilde{\mathbf{B}}_2\Delta Y_{-1})'M(\tilde{\mathbf{B}}_1\Delta Y-\tilde{\mathbf{B}}_2\Delta Y_{-1}).$$

It follows from (2.24) that

$$\hat{\sigma}^2_{v,M}-\overline{\sigma}^2_{v,M}=Q_1-EQ_1+Q_2-EQ_2-2(Q_3-EQ_3)-EQ_4, \tag{2.25}$$

where

$$Q_1 = \frac{1}{n(T-1)}\Delta Y'\tilde{\mathbf{B}}_1' M \tilde{\mathbf{B}}_1 \Delta Y,$$
$$Q_2 = \frac{1}{n(T-1)}\Delta Y_{-1}'\tilde{\mathbf{B}}_2' M \tilde{\mathbf{B}}_2 \Delta Y_{-1},$$
$$Q_3 = \frac{1}{n(T-1)}\Delta Y'\tilde{\mathbf{B}}_1' M \tilde{\mathbf{B}}_2 \Delta Y_{-1},$$
$$Q_4 = \frac{1}{n(T-1)}\left(\tilde{\mathbf{B}}_1(\Delta Y - E\Delta Y) + \tilde{\mathbf{B}}_2(\Delta Y_{-1} - E\Delta Y_{-1})\right)' P \left(\tilde{\mathbf{B}}_1(\Delta Y - E\Delta Y) + \tilde{\mathbf{B}}_2(\Delta Y_{-1} - E\Delta Y_{-1})\right).$$

Applying (2.17) with the notation $\tilde{M} = \Omega^{-1/2} M \Omega^{-1/2}$, we have

$$\begin{aligned} Q_1 =& \frac{1}{n(T-1)}(\Delta \mathbf{y}_1' \mathcal{R}' \mathbf{B}_1' \tilde{M} \mathbf{B}_1 \mathcal{R} \Delta \mathbf{y}_1 + \eta' \mathbf{B}_1' \tilde{M} \mathbf{B}_1 \eta + \Delta v' \mathbb{S}' \mathbf{B}_1' \tilde{M} \mathbf{B}_1 \mathbb{S} \Delta v \\ &+ 2\Delta \mathbf{y}_1' \mathcal{R}' \mathbf{B}_1' \tilde{M} \mathbf{B}_1 \eta + 2\Delta \mathbf{y}_1' \mathcal{R}' \mathbf{B}_1' \tilde{M} \mathbb{S} \Delta v + 2\eta' \mathbf{B}_1' \tilde{M} \mathbf{B}_1 \mathbb{S} \Delta v). \end{aligned}$$

Let $\Phi(\delta) = \mathcal{R}' \mathbf{B}_1' \tilde{M} \mathbf{B}_1 \mathcal{R}$ and $\Phi(\delta)_{ts}$ be the sub-matrix according to $t, s = 2, \ldots, T$. Q_{11}, the first term in Q_1, is written as $\frac{1}{n}\Delta y_1' \Phi_{++} \Delta y_1$ for $\Phi_{++} = \frac{1}{T-1}\sum_s \sum_t \Phi(\delta)_{ts}$. Since $\mathcal{R}, \mathbf{B}_1, \tilde{M}$ are uniformly bounded in both row and column sums by Assumptions 3, 5 and Lemma A. 3 in Yang (2018), and the elements of $\mathbf{B}_1$ are $O(h_n^{-1})$ by Assumption 5(i), Φ_{++} are uniformly bounded in either row or column sums with the elements of $O(h_n^{-1})$. By Assumption 6(iii), we have $Q_{11} - EQ_{11} = o_p(1)$ point-wise in δ. Q_{13}, the third term in Q_1, is evaluated as $\frac{1}{n(T-1)}\sum_t \sum_s \Delta v_t' \Pi_{ts} \Delta v_s$, for (t,s)th sub-matrix of $\Pi = \mathbb{S}' \mathbf{B}_1' \tilde{M} \mathbf{B}_1 \mathbb{S}$. The point-wise convergence to EQ_{13} follows from Lemma A.4(v) in Yang (2018). The 4th, 5th and 6th terms converge point-wise to the expectations in a similar manner by applying Assumption 6 (ii) and (iv), Lemma A.4 (vii) and (vi). It is shown similarly that $Q_j, j = 2, 3$ converges point-wise to the expectations.

To prove the uniform convergence, we shall prove that Q_1, Q_2, Q_3 are stochastically equicontinuous. Let l th term in Q_k be Q_{kl} and δ_1, δ_2 in Δ. By the mean value theorem,

$$Q_{kl}(\delta_2) - Q_{kl}(\delta_1) = \frac{\partial}{\partial \delta'} Q_{kl}(\overline{\delta})(\delta_2 - \delta_1),$$

for δ between δ_1 and δ_2 element-wise. Notice that Q_{kl} is linear or quadratic form of $\rho_q, \lambda_{1q}, \lambda_{2q}, q = 1, \ldots, Q$. After some algebra, we have

$$\frac{\partial}{\partial \lambda_{3q}} \tilde{M} = \tilde{M} \Omega \dot{\Omega}_q^{-1} \Omega \tilde{M},$$

where $\dot{\Omega}_q^{-1} = -C^{-1} \otimes A_3^{(q)}$, which is a linear function of $\lambda_{3q}, q = 1, \ldots, Q$. Thus, it is shown that $\sup_{\delta \in \Delta} |\frac{\partial}{\partial \delta'} Q_{kl}(\delta)| = O_p(1)$, since, for example

$$\begin{aligned} &\sup_{\delta \in \Delta} \left| \frac{1}{n(T-1)} \frac{\partial}{\partial \lambda_{3q}} \Delta \mathbf{y}_1' \mathcal{R}' \mathbf{B}_1' \tilde{M} \mathbf{B}_1 \mathcal{R} \Delta \mathbf{y}_1 \right| \\ =& \sup_{\delta \in \Delta} \frac{1}{n(T-1)} \left| \Delta \mathbf{y}_1' \mathcal{R}' \mathbf{B}_1' \tilde{M} \Omega \dot{\Omega}_q^{-1} \Omega \tilde{M} \mathbf{B}_1 \mathcal{R} \Delta \mathbf{y}_1 \right| \\ \leq& \gamma_{\max}(\dot{\Omega}_q^{-1}) \gamma_{\max}(\mathbf{B}_1' \mathbf{B}_1) \frac{1}{n(T-1)} |\Delta \mathbf{y}_1 \mathcal{R}' \mathcal{R} \Delta \mathbf{y}_1| = O_p(1), \end{aligned}$$

by Assumption 6 (i). It follows that $Q_j - EQ_j \to 0$uniformly in $\delta \in \Delta$ for $j = 1, 2, 3$.

It is left to show that $EQ_4 \to 0$ uniformly in $\delta \in \Delta$. We have

$$\begin{aligned}
EQ_4 =& \frac{1}{n(T-1)} tr\left[\Omega^{-1}\Delta X(\Delta X'\Omega^{-1}\Delta X)^{-1}\Delta X'\Omega^{-1} Var(\mathbf{B}_1\Delta Y - \mathbf{B}_2\Delta Y_{-1})\right] \\
\leq& \frac{1}{n(T-1)}\gamma_{\max}(\Omega^{-2})\gamma_{\min}^{-1}(\Delta X'\Omega^{-1}\Delta X) tr\left[\Delta X' Var(\mathbf{B}_1\Delta Y - \mathbf{B}_2\Delta Y_{-1})\Delta X\right] \\
\leq& \frac{1}{n(T-1)}\gamma_{\max}(\Omega^{-2})\gamma_{\min}^{-1}\left(\frac{\Delta X'\Omega^{-1}\Delta X}{n(T-1)}\right)\frac{1}{n(T-1)} tr\left[\Delta X' Var(\mathbf{B}_1\Delta Y - \mathbf{B}_2\Delta Y_{-1})\Delta X\right].
\end{aligned}$$

By Assumption 5(iv), $\sup_{\lambda_3\in\Lambda_3}\gamma_{\max}(\Omega^{-1}) \leq c_1 < \infty$. By Assumption 3, $\gamma_{\min}\left(\frac{\Delta X'\Omega^{-1}\Delta X}{n(T-1)}\right) \geq \inf_{\lambda_3\in\Lambda_3}\gamma_{\min}(\Omega^{-1})\gamma_{\min}\left(\frac{\Delta X'\Delta X}{n(T-1)}\right) \geq c_2 > 0$.

It follows by Assumption 3 that

$$EQ_4 \leq \frac{1}{n(T-1)} c_1^2 c_2 \frac{1}{n(T-1)} tr\left[\Delta X' Var(\mathbf{B}_1\Delta Y - \mathbf{B}_2\Delta Y_{-1})\Delta X\right] = O\left(n^{-1}\right).$$

Hence $\hat{\sigma}^2_{v,M}(\delta) - \overline{\sigma}^2_{v,M}(\delta) \to 0$ in probability uniformly in $\delta \in \Delta$.

Finally, it is seen from (2.22) and (2.24) that the right hand sides in (c), (d), (e) and (f) are expressed in the form of (2.25). Hence the arguments for (b) are applied to prove them.

2.A.2 Proof of Theorem 2

We have by the mean value theorem,

$$\begin{aligned}
0 = \frac{1}{\sqrt{n(T-1)}} S^*(\hat{\psi}_M) =& \frac{1}{\sqrt{n(T-1)}} S^*(\psi_0) \\
&+ \left[\frac{1}{n(T-1)}\frac{\partial}{\partial\psi'} S^*(\bar{\psi})\right]\sqrt{n(T-1)}(\hat{\psi}_M - \psi_0),
\end{aligned}$$

where $\bar{\psi}$ lies elementwise between $\hat{\psi}_M$ and ψ_0. The theorem will follow if

(a) $\frac{1}{\sqrt{n(T-1)}} S^*(\psi_0) \xrightarrow{D} N\left(0, \lim_{n\to\infty}\Gamma(\psi_0)\right)$,

(b) $\frac{1}{n(T-1)}\left[\frac{\partial}{\partial\psi'} S(\bar{\psi}) - \frac{\partial}{\partial\psi'} S(\psi_0)\right] \xrightarrow{p} 0$,

(c) $\frac{1}{n(T-1)}\left[\frac{\partial}{\partial\psi'} S(\psi_0) - E\left(\frac{\partial}{\partial\psi'} S(\psi_0)\right)\right] \xrightarrow{p} 0$.

Proof of (a) From (2.18), we see that $S^*(\psi_0)$ consists of three types of elements: $\Pi'\Delta v, \Delta v'\Phi\Delta v$ and $\Delta v'\Psi\Delta\mathbf{y}_1$, which can be written as $\Pi'\Delta v = \sum_{t=1}^T \Pi_t^* v_t, \Delta v'\Phi\Delta v = \sum_{t=1}^T\sum_{s=1}^T v_t'\phi_{ts}^* v_s$, and $\Delta v'\Psi\Delta\mathbf{y}_1 = \sum_{t=1}^T v_t'\Psi_t^*\Delta y_1$, where Π_t^*, Φ_{ts}^* and Ψ_t are formed by the elements of the partitioned Π, Φ and Ψ,

respectively. By (2.2), $y_1 = B_{10}^{-1}B_{20}y_0 + a + B_{10}^{-1}B_{30}^{-1}v_1$, leading to, for every non-zero $(pQ + 4Q + 1) \times 1$ vector c of constants,

$$c'S^*(\psi_0) = \sum_{t=1}^{T}\sum_{s=1}^{T} v_t'A_{ts}v_s + \sum_{t=1}^{T} v_t'B_tv_1 + \sum_{t=1}^{T} v_t'g(y_0) - c'\mu^*,$$

for suitably defined non-stochastic matrices A_{ts}, B_t, the vector μ^* and the function $g(y_0)$ linear in y_0. AS $\{y_0, v_1, \ldots, v_T\}$ are independent, the asymptotic normality follows from Lemma A5 in Yang (2018).

Proof of (b). The Hessian matrix $H(\psi) = \frac{\partial}{\partial\psi'}S(\psi)$ has elements, for $q, r = 1, \ldots, Q$,

$$\begin{aligned}
H_{\beta\beta} &= -\frac{1}{\sigma_v^2}\Delta X'\Omega^{-1}\Delta X,\\
H_{\beta\sigma_v^2} &= -\frac{1}{\sigma_v^4}\Delta X'\Omega^{-1}\Delta u(\theta),\\
H_{\beta\rho_q} &= -\frac{1}{\sigma_v^2}\Delta X'\Omega^{-1}\mathbf{W}_0^{(q)}\Delta Y_{-1},\\
H_{\beta\lambda_{1q}} &= -\frac{1}{\sigma_v^2}\Delta X'\Omega^{-1}\mathbf{W}_1^{(q)}\Delta Y,\\
H_{\beta\lambda_{2q}} &= -\frac{1}{\sigma_v^2}\Delta X'\Omega^{-1}\mathbf{W}_2^{(q)}\Delta Y_{-1},\\
H_{\beta\lambda_{3q}} &= \frac{1}{\sigma_v^2}\Delta X'\dot{\Omega}_q^{-1}\Delta u(\theta),\\
H_{\sigma_v^2\sigma_v^2} &= -\frac{1}{\sigma_v^6}\Delta u(\theta)'\Omega^{-1}\Delta u(\theta) + \frac{n(T-1)}{2\sigma_v^4},\\
H_{\sigma_v^2\rho_q} &= -\frac{1}{\sigma_v^4}(\mathbf{W}_0^{(q)}\Delta Y_{-1})'\Omega^{-1}\Delta u(\theta),\\
H_{\sigma_v^2\lambda_{1q}} &= -\frac{1}{\sigma_v^4}(\mathbf{W}_1^{(q)}\Delta Y)'\Omega^{-1}\Delta u(\theta),\\
H_{\sigma_v^2\lambda_{2q}} &= -\frac{1}{\sigma_v^4}(\mathbf{W}_2^{(q)}\Delta Y_{-1})'\Omega^{-1}\Delta u(\theta),\\
H_{\sigma^2\lambda_{3q}} &= \frac{1}{2\sigma_v^4}\Delta u(\theta)'\dot{\Omega}^{-1}\Delta u(\theta),\\
H_{\rho_q\rho_r} &= -\frac{1}{\sigma_v^2}(\mathbf{W}_0^{(q)}\Delta Y_{-1})'\Omega^{-1}(\mathbf{W}_0^{(r)}\Delta Y_{-1}) + tr(\mathbf{C}^{-1}\mathbf{D}_{-1,\rho_r}W_0^{(q)}),\\
H_{\rho_q\lambda_{1r}} &= -\frac{1}{\sigma_v^2}(\mathbf{W}_0^{(q)}\Delta Y_{-1})'\Omega^{-1}(\mathbf{W}_1^{(r)}\Delta Y) + tr(\mathbf{C}^{-1}\mathbf{D}_{-1,\lambda_{1r}}W_0^{(q)}),\\
H_{\rho_q\lambda_{2r}} &= -\frac{1}{\sigma_v^2}(\mathbf{W}_0^{(q)}\Delta Y_{-1})'\Omega^{-1}(\mathbf{W}_2^{(r)}\Delta Y_{-1}) + tr(\mathbf{C}^{-1}\mathbf{D}_{-1,\lambda_{2r}}W_0^{(q)}),\\
H_{\rho_q\lambda_{3r}} &= \frac{1}{\sigma_v^2}(\mathbf{W}_0^{(q)}\Delta Y_{-1})'\dot{\Omega}_r^{-1}\Delta u(\theta),
\end{aligned}$$

$$H_{\lambda_{1q}\lambda_{1r}} = -\frac{1}{\sigma_v^2}(\mathbf{W}_1^{(q)}\Delta Y)'\Omega^{-1}(\mathbf{W}_1^{(r)}\Delta Y) + tr(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{1r}}W_1^{(q)}),$$
$$H_{\lambda_{1q}\lambda_{2r}} = -\frac{1}{\sigma_v^2}(\mathbf{W}_1^{(q)}\Delta Y)'\Omega^{-1}(\mathbf{W}_2^{(r)}\Delta Y_{-1}) + tr(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{2r}}W_1^{(q)}),$$
$$H_{\lambda_{1q}\lambda_{3r}} = \frac{1}{\sigma_v^2}(\mathbf{W}_1^{(q)}\Delta Y)'\dot{\Omega}_r^{-1}\Delta u(\theta),$$
$$H_{\lambda_{2q}\lambda_{2r}} = -\frac{1}{\sigma_v^2}(\mathbf{W}_2^{(q)}\Delta Y_{-1})'\Omega^{-1}(\mathbf{W}_2^{(r)}\Delta Y_{-1}) + tr(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{2r}}W_2^{(q)}),$$
$$H_{\lambda_{2q}\lambda_{3r}} = \frac{1}{\sigma_v^2}(\mathbf{W}_2^{(q)}\Delta Y_{-1})'\dot{\Omega}_r^{-1}\Delta u(\theta),$$
$$H_{\lambda_{3q}\lambda_{3r}} = -\frac{1}{2\sigma_v^2}\Delta u(\theta)'\left[C^{-1}\otimes(W_3^{\prime(q)}W_3^{(r)} + W_3^{\prime(r)}W_3^{(q)})\right]\Delta u(\theta) - tr(G_3^{(q)}G_3^{(r)}),$$

where $\dot{\Omega}_q^{-1} = -C^{-1}\otimes A_3^{(q)}, \mathbf{D}_{-1,\omega} = \frac{\partial}{\partial\omega}\mathbf{D}_{-1}, \mathbf{D}_\omega = \frac{\partial}{\partial\omega}\mathbf{D}$, for $\omega = \rho_q, \lambda_{1q}, \lambda_{2q}$, $q = 1,\ldots,Q$.

It is easy to check that $\frac{1}{n(T-1)}H(\hat{\psi}_0) = O_p(1)$, $\frac{1}{n(T-1)}H(\hat{\psi}_M) = O_p(1)$, and $\frac{1}{n(T-1)}H(\bar{\psi}) = \frac{1}{n(T-1)}H(\bar{\beta},\sigma_{v0}^2,\bar{\rho},\bar{\lambda}) + o_p(1)$ by Lemma A1 of Yang (2018) and the functional form of σ_v^2 in the Hessian. Hence it is left to show that

$$\frac{1}{n(T-1)}\left[H(\bar{\beta},\sigma_{v0}^2,\bar{\rho},\bar{\lambda}) - H(\psi_0)\right] = o_p(1).$$

From

$$\begin{aligned}\Delta u(\theta) = \Delta u &- \sum_{q=1}^{Q}(\lambda_{1q}-\lambda_{1q0})W_1^{(q)} - \sum_{q=1}^{Q}(\rho_q-\rho_{q0})W_0^{(q)}\Delta Y_{-1}\\ &- \sum_{q=1}^{Q}(\lambda_{2q}-\lambda_{2q0})W_2^{(q)}\Delta Y_{-1} - \Delta X(\beta-\beta_0),\end{aligned}$$
$$\begin{aligned}\Omega^{-1}(\lambda_3) - \Omega^{-1}(\lambda_{30}) = &\sum_{q,r=1}^{Q}(\lambda_{3q}\lambda_{3r}-\lambda_{3q0}\lambda_{3r0})C^{-1}\otimes(W_3^{\prime(q)}W_3^{(r)})\\ &- \sum_{q=1}^{Q}(\lambda_{3q}-\lambda_{3q0})C^{-1}\otimes(W_3^{\prime(q)}+W_3^{(q)}),\end{aligned}$$
$$\dot{\Omega}_q^{-1}(\lambda_3) - \dot{\Omega}_q^{-1}(\lambda_{30}) = C^{-1}\otimes\sum_{r=1}^{Q}(\lambda_{3r}-\lambda_{3r0})(W_3^{\prime(q)}W_3^{(r)} + W_3^{\prime(r)}W_3^{(q)}),$$

all the stochastic elements in $\frac{1}{n(T-1)}\left[H(\bar{\beta},\sigma_{v0}^2,\bar{\rho},\bar{\lambda}) - H(\psi_0)\right]$ are linear, bilinear or quadratic in $\Delta Y, \Delta Y_{-1}$ or Δu, and linear, bilinear or quadratic in $\bar{\beta}-\beta_0, \bar{\rho}_q-\rho_{q0}$ and $\bar{\lambda}_{rq}-\lambda_{rq0}, q = 1,\ldots,Q, r = 1,2,3$. Hence they are all $o_p(1)$ by the consistency of $\hat{\psi}_M$, the equation (2.17), Lemma A1 of Yang (2018) and Assumption 6.

Finally, it is left to show that all the trace terms in

$$\frac{1}{n(T-1)}\left[H(\bar{\beta},\sigma_{v0}^2,\bar{\rho},\bar{\lambda})-H(\psi_0)\right]$$

are $o_p(1)$. Let us evaluate, e.g., the trace term in $H_{\lambda_{1q}\lambda_{1r}}$ as

$$\begin{aligned}
&\frac{1}{n(T-1)}\left[tr\left(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{1r}}(\bar{\rho},\bar{\lambda}_1,\bar{\lambda}_2)\,\mathbf{W}_1^{(q)}\right)-tr\left(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{1r}}(\rho_0,\lambda_{10},\lambda_{20})\,\mathbf{W}_1^{(q)}\right)\right]\\
&=\frac{1}{n(T-1)}\sum_{s=1}^{Q}(\bar{\rho}_s-\rho_{s0})tr(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{1r}}^{\rho_s^*}\mathbf{W}_1^{(q)})\\
&+\frac{1}{n(T-1)}\sum_{s=1}^{Q}(\bar{\lambda}_{1s}-\lambda_{1s0})tr(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{1r}}^{\lambda_{1s}^*}\mathbf{W}_1^{(q)})\\
&+\frac{1}{n(T-1)}\sum_{s=1}^{Q}(\bar{\lambda}_{2s}-\lambda_{2s0})tr(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{1r}}^{\lambda_{2s}^*}\mathbf{W}_1^{(q)}),
\end{aligned}$$

where $\mathbf{D}_{\lambda_{1r}}^{\rho_s^*}$, $\mathbf{D}_{\lambda_{1r}}^{\lambda_{1s}^*}$ and $\mathbf{D}_{\lambda_{1r}}^{\lambda_{2s}^*}$ are the partial derivatives of $\mathbf{D}_{\lambda_{1r}}$, evaluated as $(\rho_s^*,\lambda_{1s}^*,\lambda_{2s}^*)$. For simplicity, consider the case of $T=3$. Then

$$\mathbf{D}(\rho,\lambda_1,\lambda_2)=\begin{pmatrix}(B_1^{-1}B_2-2I_n)B_1^{-1} & 0\\ (I_n-B_1^{-1}B_2)^2B_1^{-1} & (B_1^{-1}B_2-2I_n)B_1^{-1}\end{pmatrix}.$$

It follows that $\mathbf{D}_{\lambda_{1r}}^{\rho_s^*}$, $\mathbf{D}_{\lambda_{1r}}^{\lambda_{1s}^*}$ and $\mathbf{D}_{\lambda_{1r}}^{\lambda_{2s}^*}$ have elements given by the multiplications of the matrices $W_1^{(r)},W_1^{(s)},W_0^{(s)},W_2^{(s)},B_1^{-1}(\lambda_1)$ and $B_2(\rho,\lambda_2)$, which are uniformly bounded in a matrix norm in the neighborhood of $(\rho_0,\lambda_{10},\lambda_{20})$ by Lemma A1 and A2 in Yang (2018). Hence $\frac{1}{n(T-1)}tr(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{1r}}^{\rho_s^*}\mathbf{W}_1^{(q)})=O_p(1)$, $\frac{1}{n(T-1)}tr(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{1r}}^{\lambda_{1s}^*}\mathbf{W}_1^{(q)})=O_p(1)$, $\frac{1}{n(T-1)}tr(\mathbf{C}^{-1}\mathbf{D}_{\lambda_{1r}}^{\lambda_{2s}^*}\mathbf{W}_1^{(q)})=O_p(1)$, leading to (b).

proof of (c). All the elements in the Hessian matrix are given by either of $v'Av-E(v'Av)$, $b'\Delta y_1-b'E(\Delta y_1)$, $\Delta y_1'F\Delta y_1-E(\Delta_y 1'F\Delta y_1)$ or $\Delta y_1'G\Delta v-E(\Delta y_1'G\Delta v)$, where the matrices A,b,F,G are all uniformly bounded in both row and columns sums by Lemma A1 of Yang (2018) and Assumption 5. The results follow by applying Lemma A4(v),(vi) of Yang (2018) and Assumption 6.

2.A.3 Proof of Theorem 3

First note that

$$\begin{aligned}
&\frac{1}{n(T-1)}\sum_{i=1}^{n}\{\hat{g}_i\hat{g}_i'-E(g_ig_i')\}\\
=&\frac{1}{n(T-1)}\sum_{i=1}^{n}\{\hat{g}_i\hat{g}_i'-g_ig_i'\}+\frac{1}{n(T-1)}\sum_{i=1}^{n}\{g_ig_i'-E(g_ig_i')\}.
\end{aligned}$$

The first term is $o_p(1)$ by the mean value theorem with the arguments we used in the proof of Theorem 2(b). For the second term, the matrices to compose $g_i, i = 1, 2, 3$ in (2.19) are $\Pi_j^{(q)}, \Phi_j^{(q)}$ and $\Psi_j^{(q)}$ for $q = 1, \ldots, Q$, whose elements are all uniformly bounded by Lemma A1 in Yang (2018) under Assumptions 3 and 5. It follows that the arguments in the proof of Theorem 3.3 in Yang in (2018) can be applied to prove that the second term is $o_p(1)$.

2.A.4 Simulation results of submodels

	TRUE	Mean	sd	Bias	RMSE	$\hat{se}$	$s\hat{e}_H$
λ_{11}	0.5	0.4889	0.0599	-0.0111	0.0607	0.0708	0.0622
ρ_1	0.2	0.1991	0.0317	-0.0009	0.0316	0.0302	0.0293
β_1	7	7.0117	0.2440	0.0117	0.2431	0.2196	0.2201
λ_{12}	0.2	0.2052	0.0621	0.0052	0.0620	0.0618	0.0598
ρ_2	0.5	0.4934	0.0676	-0.0066	0.0675	0.0704	0.0682
β_2	3	3.0026	0.2456	0.0026	0.2444	0.1988	0.2076
σ^2	1	0.9674	0.1216	-0.0326	0.1253	0.1121	0.1148

Table 2.12: Simulation result of M-estimator, SL submodel

	TRUE	Mean	sd	Bias	RMSE	$\hat{se}$	$s\hat{e}_H$
λ_{31}	0.5	0.4923	0.1618	-0.0077	0.1612	0.1611	0.1625
ρ_1	0.2	0.1951	0.0314	-0.0049	0.0316	0.0304	0.0301
β_1	7	6.9776	0.2096	-0.0224	0.2098	0.2006	0.2056
λ_{32}	0.2	0.2082	0.1661	0.0082	0.1654	0.1630	0.1638
ρ_2	0.5	0.4974	0.0710	-0.0026	0.0707	0.0768	0.0692
β_2	3	2.9769	0.2127	-0.0231	0.2129	0.2108	0.2105
σ^2	1	0.9435	0.1083	-0.0565	0.1217	0.1123	0.1139

Table 2.13: Simulation result of M-estimator, SE submodel

	TRUE	Mean	sd	Bias	RMSE	$\hat{se}$	$s\hat{e}_H$
λ_{11}	0.5	0.4730	0.0686	-0.0270	0.0734	0.0870	0.0668
λ_{21}	0.1	0.1120	0.0765	0.0120	0.0770	0.1236	0.0718
ρ_1	0.2	0.1980	0.0363	-0.0020	0.0362	0.0324	0.0326
β_1	7	6.9831	0.2653	-0.0169	0.2645	0.2238	0.2255
λ_{12}	0.1	0.1124	0.0642	0.0124	0.0650	0.0709	0.0649
λ_{22}	0.2	0.2043	0.0620	0.0043	0.0619	0.0802	0.0670
ρ_2	0.5	0.4810	0.0753	-0.0190	0.0773	0.0713	0.0691
β_2	3	2.9771	0.2281	-0.0229	0.2281	0.2088	0.2138
σ^2	1	0.9614	0.1121	-0.0386	0.1180	0.1156	0.1146

Table 2.14: Simulation result of M-estimator, SLTL submodel

Bibliography

Aquaro, M., Bailey, N., and Pesaran, M. H. (2015). Quasi maximum likelihood estimation of spatial models with heterogeneous coefficients. *USC-INET Research Paper*, (15-17).

Baltagi, B. H., Song, S. H., and Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of econometrics*, 117(1):123–150.

Elhorst, J. P. (2014). *Spatial econometrics: from cross-sectional data to spatial panels*, volume 479. Springer.

Fallahi, F. (2011). Causal relationship between energy consumption (ec) and gdp: a markov-switching (ms) causality. *Energy*, 36(7):4165–4170.

Hansen, B. E. (1999). Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of econometrics*, 93(2):345–368.

Kapoor, M., Kelejian, H. H., and Prucha, I. R. (2007). Panel data models with spatially correlated error components. *Journal of econometrics*, 140(1):97–130.

Konen, W. and Hansen, N. (2015). *R-to-Java Interface for 'CMA-ES'*.

Lee, L.-f. and Yu, J. (2010). Some recent developments in spatial panel data models. *Regional Science and Urban Economics*, 40(5):255–271.

LeSage, J. P. and Chih, Y.-Y. (2018). A bayesian spatial panel model with heterogeneous coefficients. *Regional Science and Urban Economics*, 72:58–73.

Mahalingam, B. and Orman, W. H. (2018). Gdp and energy consumption: A panel analysis of the us. *Applied Energy*, 213:208–218.

Majumdar, A., Gelfand, A. E., and Banerjee, S. (2005). Spatio-temporal change-point modeling. *Journal of Statistical Planning and Inference*, 130(1-2):149–166.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tong, H. and Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data-with discussion. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 42(3):245–292.

Yang, Z. (2018). Unified m-estimation of fixed-effects spatial dynamic models with short panels. *Journal of econometrics*, 205(2):423–447.

Yu, J., De Jong, R., and Lee, L.-f. (2008). Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and t are large. *Journal of Econometrics*, 146(1):118–134.

Chapter 3

Estimating spatial regression models with sample data points: a Gibbs sampler solution

3.1 Introduction

In the spatial analysis of individual data (Arbia et al., ming), often the observations employed in the estimation of statistical models do not represent the full population, but only a sample from it. Furthermore, in many instances, samples do not even obey any specific design and they are collected only with a convenience criterion as it happens e. g. when data are webscraped or crowdsourced (Arbia and Nardelli, 2020; Cavallo and Rigobon, 2016). In the literature on point pattern analysis it is common to distinguish the situation when all points are available (called a mapped pattern) from the situation when only a sample of them can be observed (referred to as a sample pattern. See Diggle (1983); Illian et al. (2008)). Although the case of sample pattern is extremely common in practical cases, it has been rather overlooked so far in the econometric literature on spatial regression which generally just assume the problem away and treats the collection of data as if they constitute the entire population (see e. g. Arbia (2011)). A somewhat related problem has been approached in the literature when considering the case of missing spatial data. In this area of research, in a series of papers in the '80s, Bennett, Griffith, and Haining analyzed the effects of missing spatial data and compared the performances of different methods to replace them (Bennett et al., 1984; Haining et al., 1984; Griffith et al., 1989). More recently Arbia et al. (2016) have reappraised the problem extending the study to the effects on the estimation of a spatial regression model. They show that the presence of missing data reduces the precision of the estimates of all the

regression parameters with a reduction of the efficiency which is emphasized by the presence of strong spatial correlation and by the presence of missing points which are clustered in space (Arbia et al., 2016). The problem of missing data is well known in the statistical literature (Little, 1988; Little and Rubin, 2019; Rubin, 1976) where solutions have been suggested to replace the observations that are missing following different interpolation strategies (Dempster et al., 1977; Rubin, 2004), although with no explicit reference to the spatial data peculiarities. The problem of analyzing data with relevant spatial characteristics missing at random has been also addressed in the econometric literature by Baltagi et al. (2007); Kelejian and Prucha (2010); Flores-Lagunes and Schnier (2012); Muris (2020). Although related to the problem of spatial samples, the issue of missing spatial data differs for at least a couple of reasons. First of all, when data are missing we can expect that most of the data are observed and that only a relatively small proportion of them is lost. In the situation we have in mind it is exactly the opposite: when data are sampled, only a small proportion of the population data are observed and most of the others remain unobservable. Secondly, in the first case missing data usually do not follow any precise scheme, while in the case which is of interest for us, in some cases we can assume that the sample design can be taken under control. The paper is organized as follows. In Section 2, we present the results of a set of Monte Carlo experiments, to examine the consequences of the estimation of a spatial econometric (Anselin, 1988; Arbia, 2014). In Section 3 we propose the use of the Gibbs sampler to replace the unobserved data points thus reducing the distorting effects induced by the sample observation of a wider phenomenon. We also report the results of a Monte Carlo experiment to assess the performances of the proposed method while in Section 4 we evaluate them through a case study based on some land price data in Tokyo, Japan. Section 5 concludes.

3.2 A Monte Carlo evaluation of the effects of sampling on the estimation of the Spatial Lag Model parameters

In this section, we aim at shedding light on the possible consequences of observing a sample of data when the observations are distributed in space following a certain point pattern. In particular, through a set of Monte Carlo experiments, we will examine what is the consequence of the ML estimation of the parameters of a spatial regression model. Our goal is to isolate three relevant effects, namely:

(i) the effect of the sample size,

(ii) the effects of the pattern distribution of the data points at the population level,

(iii) the effects of the sampling criterion.

These three experimental situations will be presented in turn in the next sections. In all simulations we will consider, in particular, the following Spatial Lag Model (henceforth SLM, see Arbia (2014)):

$$y = \beta x + \rho W y + \sigma\varepsilon \tag{3.1}$$

where y is an $n \times 1$ vector of observations of the dependent variable, x is an $n \times 1$ vector of observations of the independent variable (for the sake of simplicity we consider only one predictor in our model), W is an exogenously specified weight matrix which takes care of the links of proximity between the n units and $\varepsilon \sim i.i.d.N(0, 1)$ are standardized normal independent innovations.

3.2.1 Assessing the effects of the sample proportion

To start with, in order to isolate the effects of the sample proportion, in our first experiments we considered Model (3.1) in the specific case when W is specified as an inverse squared distance, the model's parameters are set to $\beta = 1$ and $\rho = 0.5$ and with error variance $\sigma^2 = 1$.Furthermore, we will consider five different populations sizes $N = [500, 750, 1000, 2000, 3000]$, with the points that are randomly distributed according to a Complete Spatial Randomness scheme (CSR, Diggle, 1983) in a unitary square so that both coordinates will be generated by two independent uniform distributions between 0 and 1. Finally, point sample data are selected with the simple random sample procedure.

First of all, we simulate the vector of observations of the exogenous regressor from a Gaussian distribution with unitary expected value and unitary variance. Furthermore, in each simulation run, the errors ε of Equation (3.1) are generated from a standardized normal distribution. In each replication of the Monte Carlo experiments, the vector y is then generated through the following expression:

$$y = (I - \rho W)^{-1}\beta x + \sigma^2(I - \rho W)^{-1}\varepsilon \tag{3.2}$$

All experiments are replicated 1,000 times.

In each replication we then estimate the parameters β and ρ with a Maximum Likelihood procedure both observing the entire population and in a sample experiment by considering a sample proportion ranging from 2% to 20%, $\frac{n}{N} = [0.02, 0.04, 0.06, 0.08, 0.10, 0, 0.12, 0.14, 0.16, 0.18, 0.20]$.

We start considering the effect of sampling data on the estimation of the regression parameter β. Figure 3.1 shows the box-plots of the simulated sampling distribution of the regression coefficient β when we use the whole population compared with those obtained when we consider only a sample with a sample proportion of 2%.

The figure shows that the estimation of the regression parameter β is approximately unbiased if we use all population values (Figure 3.1a). In contrast, a positive bias is present if we use only sample values (Figure 3.1b) with the additional feature of a bias reduction when the population size N increases.

A second remarkable feature is the larger standard errors displayed by the sampling distribution when using only sample data. We observe this effect at all

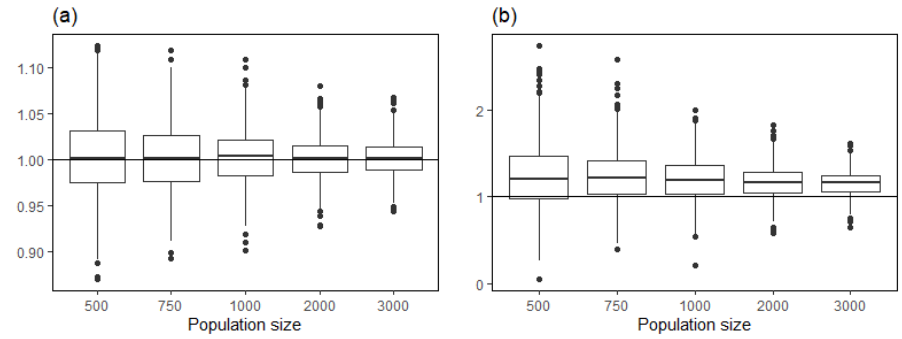

Figure 3.1: Simulated sampling distribution of the regression coefficient β (in Equation (3.1)) using all population values (a) and sample values (b) with a sample proportion $\frac{n}{N} = 0.02$. In each graph we consider different population dimensions ($N = [500, 750, 1000, 2000, 3000]$). The true value of β is 1, indicated by the horizontal line.

population dimensions (compare Figure 3.1a with Figure 3.1b. Notice that two different scales for the vertical axis are used in the two graphs).

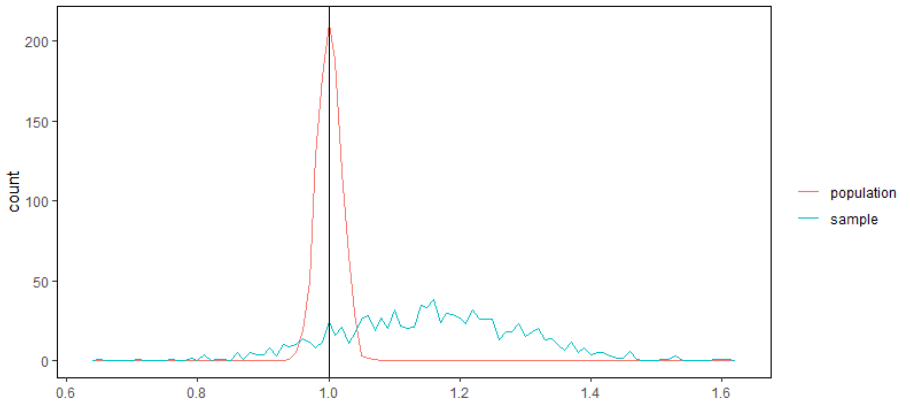

Figure 3.2: Simulated sampling distribution of the regression coefficient β using all population values (red lines) and sample values (blue line) with a population dimensions $N = 3000$ and a sample proportion $\frac{n}{N} = 0.02$. The true value of β is represented with a vertical line.

For the sake of illustration, two sampling distributions are reported on the same graph in Figure 3.2 in the specific case of $N = 3000$ and $\frac{n}{N} = 0.02$. Figure 3.2 clearly shows the larger bias and the greater inefficiency of the estimation of β when considering sample point data.

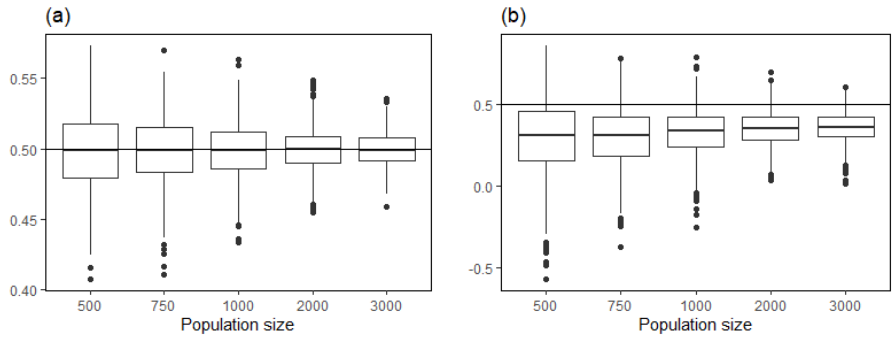

Figure 3.3: Simulated sampling distribution of the regression coefficient ρ using all population values (a) and sample values (b) with a sample proportion $\frac{n}{N} = 0.02$. We consider different population dimensions $N = [500, 750, 1000, 2000, 3000]$. The true value of ρ is 0.5, and is indicated by the horizontal line.

Let us now consider the effects of sample observations in the estimation of the spatial correlation parameter ρ. Figure 3.3a shows a small downward bias if we use all population values (as established by Smith (2009)). Note that Smith's result holds specifically in the case of strongly connected weight matrices which is exactly our case. If, however, we use sample values instead of the whole population we observe a much larger bias (See Figure 3.3b).

Similarly, to what we observed for the estimation β, also in the case of estimating ρ we have much larger standard errors when we use only sample data whatever is the population size (compare Figure 3.3a with Figure 3.3b).

Again, only for the sake of illustration, in Figure 3.4 we report on the same graph the two sampling distributions obtained with population and with sample data in the case of $N = 3000$. Similarly to what we observed commenting on Figure 3.2, Figure 3.4 also shows a much larger bias and a higher inefficiency in the estimation of ρ when we consider sample data.

Finally, Figure 3.5 reports the Monte Carlo evaluation of the standard error of the sampling distribution of the regression coefficient β and of the spatial correlation coefficient ρ when using sample values at different sample proportions increasing from 2% to 20%. Both graphs reported in Figure 3.5 clearly show the expected decrease in the standard errors (so the increase in efficiency) of the estimators of β and ρ when the sample proportion increases.

In summary, our Monte Carlo experiments show that, in the presence of sample data, we observe biased and inefficient estimates for both β and ρ.

3.2.2 Assessing the effects of the data point pattern

Let us now move to consider the effects on the reliability of estimates based on sample data due to the pattern of the point data at the population level. Indeed, so far we have considered the case when the population data-points

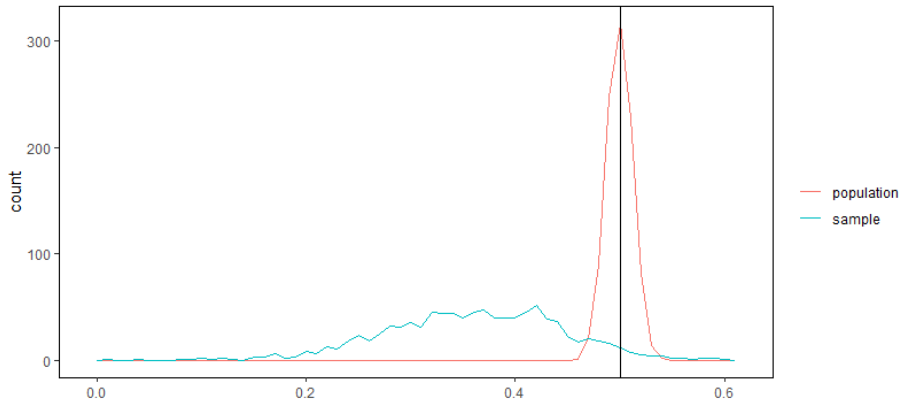

Figure 3.4: Simulated sampling distribution of the regression coefficient ρ using all population values (red lines) and sample values (blue line) with a population dimensions $N = 3000$ and a sample proportion $\frac{n}{N} = 0.02$. The true value of β is represented with a vertical line.

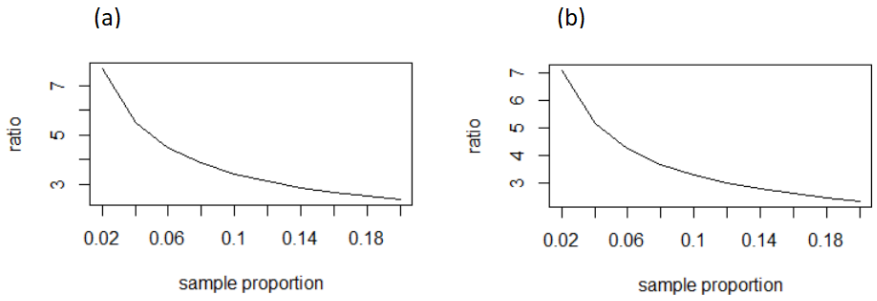

Figure 3.5: Monte Carlo evaluation of the standard error of the sampling distribution of the regression coefficient β (a) and of the spatial correlation coefficient ρ (b) using sample values at different sample proportions ranging from 2% to 20%.

are scattered in space according to a CSR pattern and the observed points are randomly selected.

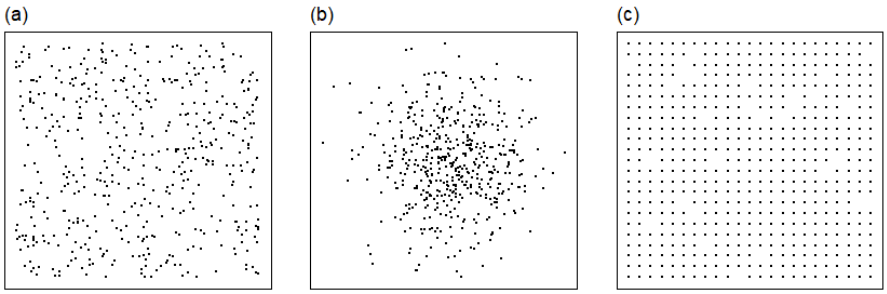

Figure 3.6: (a) Random, (b) Clustered, and (c) Inhibitory point patterns

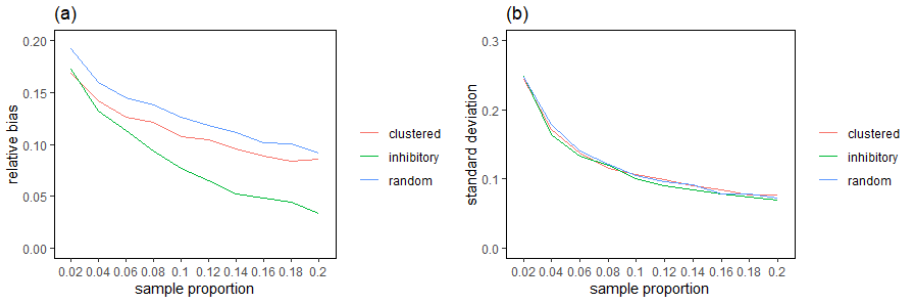

Figure 3.7: Plot of the relative bias (a) and standard deviation (b) of estimated β at different sample proportion, $N = 1000$. Mean of 1000 iterations.

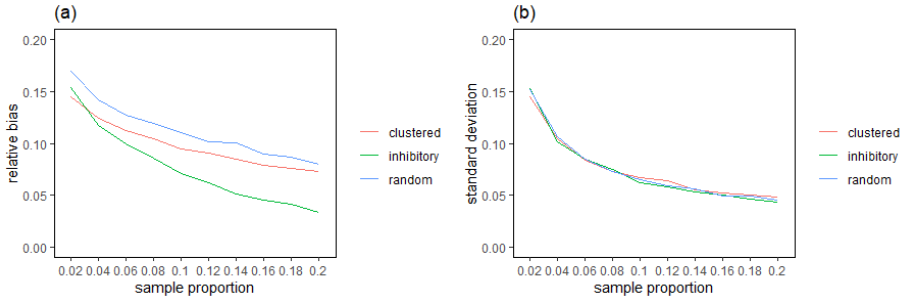

Figure 3.8: Plot of the relative bias (a) and standard deviation (b) of estimated ρ at different sample proportion, $N = 1000$. Means of 1000 iterations.

In this second Monte Carlo experiment, we aim at extending our findings to a larger set of situations. In particular, we will consider three different patterns

for the population data points. We start examining again, for comparison, the case of points which are randomly distributed at the population level, obeying a CSR. (Figure 3.6a). We will then consider the case when data-points are clustered at the population level (Diggle, 1983) leading to an extra concentration in the center of the unitary square. In this case, we will consider the points generated by a truncated bivariate normal distribution of the coordinates with a mean of 0.5 and a standard deviation of 0.15 in both directions. (Figure 3.6b). Finally, we will consider the case when points follow an inhibitory pattern at the population level (Diggle, 1983), and thus they are regularly distributed in the unitary square (Figure 3.6c). In all three cases, the sample data points are randomly selected as in Section 3.2.1. Results are reported in Figures 3.7 for the slope coefficient and in Figures 3.8 for the correlation parameter.

The two figures display a very similar behavior and they highlight two remarkable features shared by the two parameters. First of all, in terms of the bias, in all three scenarios, the estimations of both the regression slope and of the spatial correlation parameter converge to the true value with a relative bias that decreases monotonically when the sample proportion increases. However, the presence of a form of non-randomness in the point pattern (either in the form of cluster or inhibition) reduces the relative bias at each sample proportion. The lowest bias is achieved when the point pattern displays inhibition. This effect can be rationalized as follows. When data are positively spatially correlated points that are close-by tend to be similar. In this situation, if points are originally clustered or regularly distributed, a random sampling tends to cancel this feature (at least in small samples) thus increasing the bias in the estimators. The second feature which is rather evident looking at Figures 3.7 and 3.8, is that the pattern of the points seems to affect only marginally the efficiency of the estimators.

3.2.3 Assessing the effects of the sampling criterion

Let us now move to consider the effects on the reliability of estimates based on sample data due to the sampling criterion used in the selection. To this aim, in this third Monte Carlo experiment, we will consider four different sampling criteria.

The first design (Design a) represents the benchmark where the data points are randomly selected from the population. In the other three designs (Design b, c, and d) we try to mimic the situation occurring when data are collected without a proper sample design and following only a convenience criterion as it happens, e. g. when data are collected through webscraping or crowdsourcing. In particular, in Design b we divide the population points into four quadrants, and we sample from only one of the four quadrants randomly chosen. This solution mimics the case of geographical selection bias. We call this design, quadrant sampling. In Design c, we divide the population points into a 10×10 grid, we randomly pick up some cells of the grid, we use the points in the grid as sample units and we repeat sequentially the operation until reaching the target sample size. This situation mimics the case of Contagion which we observe in many crowdsourced surveys (see Arbia et al. (2020)). Finally, in Design d we

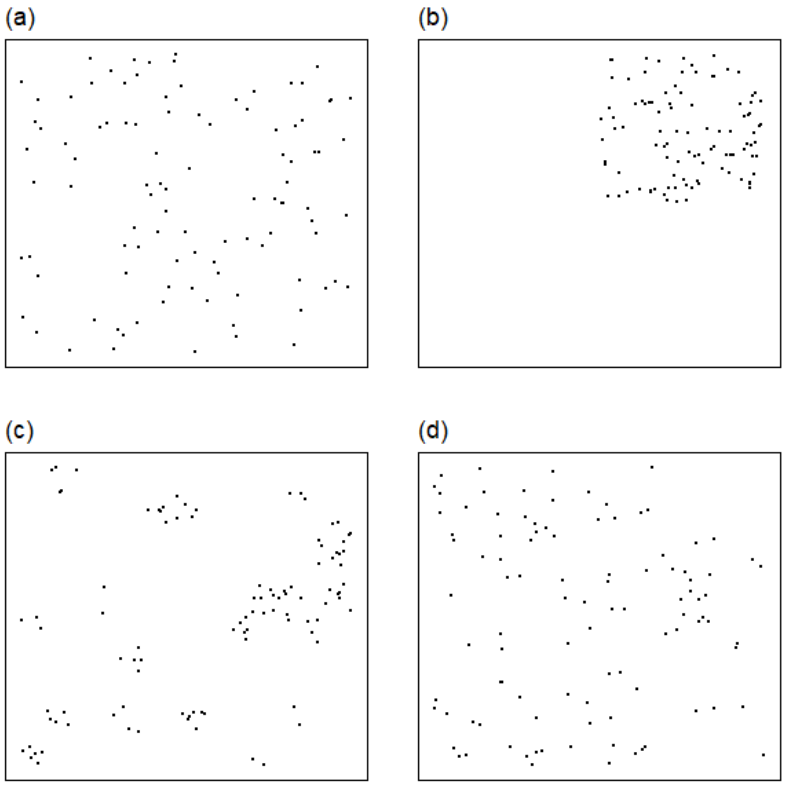

Figure 3.9: Illustration of four different sampling designs (N = 500, sample size = 100). (a) simple random, (b) quadrant, (c) contagion, (d) threshold.

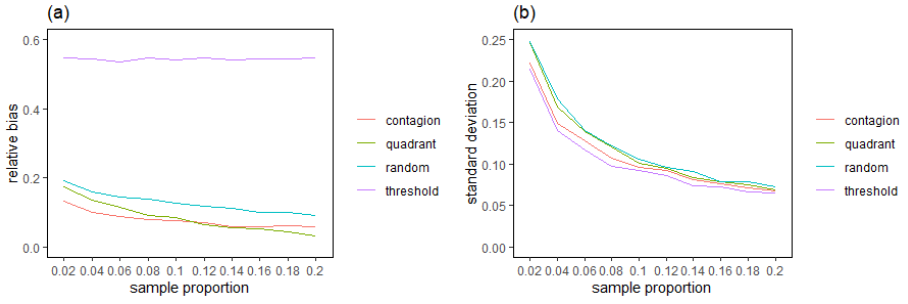

Figure 3.10: Relative bias (a) and standard deviation (b) of the estimation of β at different sample proportion with different sampling designs. N = 1000. Means of 1000 iterations.

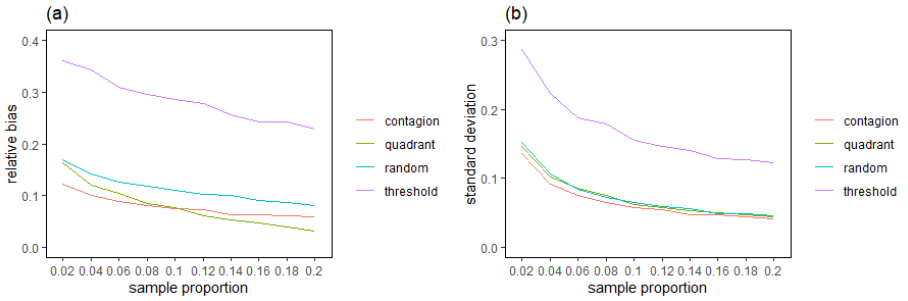

Figure 3.11: Relative bias (a) and standard deviation (b) of the estimation of ρ at different sample proportions with different sampling designs. Means of 1000 iterations.

draw a random sample from the point pattern only if it presents a value of the dependent variable lower than the 30th percentile. This case mimics situations where data are self-selected on the basis of the value of the observed variable like it happens sometimes with webscraped data (Arbia and Nardelli, 2020). We will refer to this fourth case as the Threshold design. A graphical representation of the four sample designs is reported in Figure 3.9. The results of our simulations are reported in Figures 3.10 and 3.11 respectively for the two spatial regression parameters.

The graphs reported in Figures 3.10 and 3.11 display very similar behaviors for both parameters with respect to the estimation bias. In particular, in the threshold sample, the relative biases of both parameters are very high and consistently much larger than those obtained with the other three sampling schemes. The lowest bias is achieved when data are collected with a contagion scheme at small sampling proportions although when the sample size increases the quadrant scheme produces more reliable estimates.

Contrasting results are obtained when analyzing the effects of sampling on the efficiency of the estimators. In this case, in fact, the results produced with the four schemes are very similar in the case of the regression slope, while in the case of the correlation parameter the threshold data collection is the one that produces remarkably higher variability with respect to the other designs.

In summary, if the convenience sampling follows a threshold criterion, we obtain the largest bias in the estimation of β, and the highest error variance in the case of the correlation coefficient. Quadrant and contagion, generally speaking, mitigate both the bias and the loss in efficiency.

3.3 A Gibbs sampler solution in the estimation of a spatial regression based on sample point data

In the previous section we have shown, through a series of MC experiments, that if we estimate a spatial regression model using data sampled from a larger population of points, the regression parameters will not be estimated accurately, in terms of both bias and inefficiency. In this section, we propose a solution to mitigate such inaccuracies.

Let us consider the case where we observed n sample units of the variable y, say $y_n = (y_1, \ldots, y_n)$, drawn from a finite population consists of N points, $N > n$, so that $y_N = (y_1, \ldots, y_N)$, and that the value of y_N at the population level is generated by the SLM reported in Equation (3.1) and repeated here in the form of a vector of independent variables:

$$y_N = \rho W_N y_N + X_N \beta + \sigma \varepsilon_N \tag{3.3}$$

In Equation (3.3), W_N represents an $N \times N$ row normalized spatial weight matrix, X_N is a $N \times p$ design matrix of independent variables and ε_N is a standardized normal random vector with zero mean and unitary covariance matrix, say I_N. Let us consider an estimation for the parameters ρ, β, and σ obtained with the available subset of observations y_n, assuming that the design matrix X_N is known without error. In this section, we also consider the case when the exact position of both in and out of sample points are known so that the population spatial weight matrix W_N is also known without error. First of all, let us write the subset model based on sample data as follows:

$$y_n = \rho W_n y_n + X_n \beta + \sigma \varepsilon_n \tag{3.4}$$

where W_n is the row normalized submatrix of W_N corresponding to the subset of the observed locations.

we propose a Bayesian approach to reduce the bias in the result of the maximum likelihood approach observed in the previous section, when only sample data is used for estimation. In particular, we will assume a Gaussian prior for ρ with an expected value corresponding to the ML estimator (denoted by $\hat{\rho}_{MLE}$) and a variance provided by the negative inverse of the Hessian associated with the likelihood function (denoted by $\hat{\tau}_{MLE}$). We will further assume non-informative priors for the other model's parameters. Then a Gibbs sampling procedure can be conducted as follows:

(i) Initialize the process by setting $y_N = (y_n, 0_{N-n})$ and $\rho = 0$.

(ii) Draw a posterior sample of ρ conditional on y_N using the concentrated likelihood.

(iii) Draw the posteriors samples of β and σ conditional on y_N and ρ.

(iv) Build up posterior samples of $y_{n+1}, \ldots, y_N$ by $y_j = \rho W_{Nj} y + X_j \beta$ conditional on ρ and β, where W_{Nj} the jth row of W_N.

(v) Draw $\frac{1}{\sigma^2}$ from a Gamma distribution, $Gamma(\frac{n}{2}, \frac{(y_n - \rho W_n y_n - X_n\beta)'(y_n - \rho W_n y_n - X_n\beta)}{2})$, conditional on ρ and β.

(vi) Return to Step 2 and iterate.

We would like to clarify some details about the procedure. Firstly, in Step (ii), since no explicit form of the posterior distribution is available, we draw the sample from the posterior approximated by a Normal distribution. Specifically, we approximate the posterior by a Normal distribution with the expected value obtained by maximizing the following expression:

$$L(\rho) = -\frac{N}{2}\log\hat{\sigma}^2 + \log||I_N - \rho W_N|| - \frac{1}{2}\log\hat{\tau}_{MLE} - \frac{(\rho - \hat{\rho}_{MLE})^2}{2\hat{\tau}_{MLE}} \tag{3.5}$$

with the term $\hat{\tau}_{MLE}$ given by the inverse of the negative Hessian at the maximizing point, where

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{e_N' e_N}{N} \\
e_N &= y_N - \hat{\rho} W_N y_N - X_N \hat{\beta} \\
\hat{\beta}_N &= (X_N' X_N)^{-1} X_N' (y_N - \rho W_N y_N)
\end{aligned}$$

and e is a vector of empirical residuals. Secondly, in Step (iii), the posteriors are explicitly obtained as the Gaussian distribution by the standard formula of linear regression models. Thirdly, in Step (iv), m iterations are considered to validate the convergence of the recursion to construct the out-of-sample value estimators. Notice that the error term, which should be included in normal Gibbs sampling procedures, is not inserted in the recursion. This allows us to mitigate the negative bias of ρ at the price of the negative bias for σ^2 in Step (iii), a problem that will be solved in the next step. Finally, in Step (v), we re-estimate σ^2 using only available samples in order to avoid the negative bias caused in Step (iii).

To test the performances of the proposed procedure, we conducted the following simulation study. First of all, we simulate a point pattern of 500 points following a CSR at the population level and a 500×1 vector of observations of the independent variable x from i.i.d. normal distribution with an expected value equal to 1 and unitary variance. Then we compute the weight matrix W using the inverse squared distance and we generate the observations of the dependent variable y using the SLM specification (see Equation (3.1)) with $\rho = 0.5, \beta = 5, \sigma^2 = 1$ and with an intercept equal to 1. Finally, we select 10% of the observations as the observed sample ($n = 50$), employing the four methods illustrated in the previous section, namely: (i) random; (ii) quadrants, (iii) contagion and (iv) threshold. Each experiment is replicated 100 times. For the Gibbs sampler in our proposed procedure, we iterate 4,000 times and use

	population	sample	
1) Random	MLE	MLE	Gibbs
ρ_{mean}	0.497	0.127	0.425
ρ_{RMSE}	0.018	0.390	0.105
$intercept_{mean}$	1.020	5.260	1.879
$intercept_{RMSE}$	0.218	4.495	1.208
β_{mean}	5.009	5.196	5.020
β_{RMSE}	0.045	0.302	0.168
σ^2_{mean}	1.004	2.315	3.013
σ^2_{RMSE}	0.066	1.527	2.337

	population	sample	
2) Quadrant	MLE	MLE	Gibbs
ρ_{mean}	0.497	0.255	0.414
ρ_{RMSE}	0.017	0.264	0.112
$intercept_{mean}$	1.033	3.867	2.045
$intercept_{RMSE}$	0.220	3.126	1.364
β_{mean}	4.999	5.099	5.002
β_{RMSE}	0.048	0.251	0.164
σ^2_{mean}	1.003	1.724	1.961
σ^2_{RMSE}	0.070	0.871	1.130

	population	sample	
3) Contagion	MLE	MLE	Gibbs
ρ_{mean}	0.499	0.278	0.393
ρ_{RMSE}	0.017	0.229	0.130
$intercept_{mean}$	1.012	3.563	2.235
$intercept_{RMSE}$	0.215	2.681	1.526
β_{mean}	5.003	5.090	5.049
β_{RMSE}	0.052	0.199	0.170
σ^2_{mean}	0.993	1.364	1.592
σ^2_{RMSE}	0.065	0.510	0.804

	population	sample	
4) Threshold	MLE	MLE	Gibbs
ρ_{mean}	0.498	0.115	0.447
ρ_{RMSE}	0.018	0.421	0.091
$intercept_{mean}$	1.038	5.607	1.676
$intercept_{RMSE}$	0.220	4.732	1.030
β_{mean}	5.003	4.170	4.487
β_{RMSE}	0.042	0.937	0.635
σ^2_{mean}	0.990	1.986	6.685
σ^2_{RMSE}	0.060	1.131	6.038

Table 3.1: Mean and RMSE for a 10% sample ($n = 50$) drawn from the population points ($N = 500$) evaluated by 100 simulations. Comparisons between the MLEs based on the population, the sample data and sample data corrected with the Gibbs sampling procedure in the four sampling schemes.

the latter 3,000 as the posterior sample. The mean values and the Root Mean Squared Errors (RMSE) obtained in the simulation exercise are reported in Table 3.1.

Table 3.1 shows that in accordance with the findings of Section 3.2, the Maximum Likelihood estimation of the parameter ρ is subject to a strong underestimation especially if the sample is selected randomly or with the threshold method. In contrast, our proposed method achieves better results, consistently producing lower biases and lower standard errors than those obtained with a Maximum Likelihood procedure. Similar results are observed when considering the estimation of the intercept and of the regression slope. On the other hand, our method introduces a larger bias in the estimation of σ^2 especially in the case of the threshold sample, a result of using ρ and β estimated from the population which does not minimize the error for the sample data set.

3.4 A case study: hedonic land price modeling in western Tokyo

In this section, we aim at showing the practical advantages of our proposed estimation method, using the land price data in western Tokyo. Land price data at 513 observation points in the western area of Tokyo are collected by the Japanese Government in 2019 and publicly available. The geographical locations of the data points are reported in Figure 3.12.

At each point, we collected the data related to the land price per square meters together with the geographical coordinates expressed in longitude and

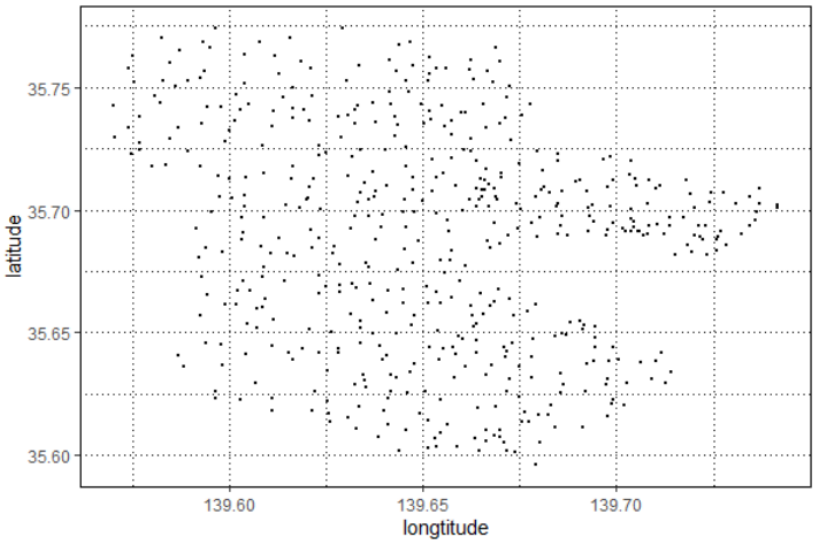

Figure 3.12: Location of 513 points of land price data in western Tokyo,2019.

latitude. The land price data are transformed into a growth ratio in the period 2018-2019 defining our dependent variable in terms of the log difference:

$$y_j = \log P_{j,2019} - \log P_{j,2019}, j = 1, \ldots, 513. \tag{3.6}$$

We consider two independent variables from the same source, namely: the log-distance from the Shinjuku terminal station and the log-distance from the closest station. Then We randomly selected 150 sample observations from the complete dataset (corresponding to 29% of the population points) and then estimate and compare the results obtained with the following three mothods:

(i) The Maximum Likelihood method using the complete dataset,

(ii) The Maximum Likelihood method using only the 150 sample observations,

(iii) Our proposed method based on the Gibbs sampling correction strategy using 150 observations.

In order to eliminate the randomness in the choice of the samples, we repeat method (ii) and (iii) 100 times and take the average results. When using the Gibbs sampling strategy, we iterate the Gibbs procedure 20000 times, and obtain the estimation results using the average of every 5th posterior samples (see Section 3.3).

The estimation results obtained with the three methods are reported in Table 3.2. β_1 corresponds to log-distance from the Shinjuku terminal station and β_2 to the closest station. In particular, Table 3.2 shows that the application of the Gibbs sampling correction procedure succeeds in reducing the negative

	MLE population		MLE sample			Gibbs sampler correction			
	estimation	se	estimation	relative bias	se	estimation	relative bias	se	relative efficiency
	(1)	(2)	(3)	(4) = $\frac{((3)-(1))}{(1)}$	(5)	(6)	(7)= $\frac{((6)-(1))}{(1)}$	(8)	(9)= $\frac{(8)}{(5)}$
ρ	0.944	0.055	0.663	-0.298	0.206	0.741	-0.215	0.251	1.218
Intercept	6.725	0.648	9.058	0.347	2.580	8.443	0.255	2.992	1.160
β_1	-0.345	0.060	-0.419	0.214	0.154	-0.399	0.157	0.165	1.071
β_2	-0.480	0.062	-0.498	0.038	0.138	-0.498	0.038	0.138	1.000
σ^2	3.416	0.214	3.499	0.024	0.625	3.488	0.021	0.626	1.000

Table 3.2: Comparisons of MLE result of whole dataset, MLE and Gibbs sampler result of samples, mean and standard errors from 100 iterations for the latter two estimators.

bias in the estimation of the spatial correlation parameter and of all regression parameters. It also reports systematically lower standard error of the estimators for the Gibbs sampler correction with respect to the ML-based on incomplete data (see column (9) in Table 3.2).

3.5 Conclusions

The aim of this paper was to draw the researchers' attention in spatial econometric studies, on the inappropriate use of spatial data collected through some point sampling collection and, in particular, when adopting some of the unconventional collection methods currently available, e. g., data webscraping or crowdsourcing. In particular, the first part of the paper is devoted to shedding light on the biases and inefficiencies associated with the estimation of spatial regression parameters when the research does not have the availability of the whole set of data points, but only of a subset of them. Our simulation studies show that the estimators of the regression coefficients and of the spatial correlation parameter in a spatial regression are affected by higher bias and by a lower efficiency if they are based on sample data. Starting from these results, in the second part of the paper, we suggest a possible solution to mitigate such distorting effects. In this respect, we propose a Gibbs sampler solution to replace the unsampled observation prior to the estimation phase. Through a series of Monte Carlo studies, we show that this approach generally achieves lower bias and lower standard errors in the estimation of the parameters of a Spatial Lag model. The field of application of this solution is limited by the fact that we were forced to assume a full knowledge about the spatial coordinates of all the population points (including those that are unsampled) and, consequently, about the weight matrix. There are indeed many empirical cases where such an assumption can be realistic and a full list of all the population elements is available. This happens, for instance, in industrial economic studies where firm locations are available, but some variables can only be observed through sample surveys. However, it is fair to remark that in many other cases both the location and the value of some variables remain unknown to the researcher like e. g. in many spatial sample surveys, or when we scrape the data from the web. For this second interesting case, the method proposed

here cannot be applied and ad hoc solutions to mitigate the distortions are left outside the scope of the present paper and will be considered in some future work.

Bibliography

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, volume 4. Springer Science & Business Media.

Arbia, G. (2011). A lustrum of sea: recent research trends following the creation of the spatial econometrics association (2007–2011). *Spatial Economic Analysis*, 6(4):377–395.

Arbia, G. (2014). *A primer for spatial econometrics with applications in R*. Springer.

Arbia, G., Espa, G., and Giuliani, D. (2016). Dirty spatial econometrics. *The Annals of Regional Science*, 56(1):177–189.

Arbia, G., Espa, G., and Giuliani, D. (forthcoming). *Spatial microeconometrics*. Routlegde.

Arbia, G. and Nardelli, V. (2020). On spatial lag models estimated using crowdsourcing, web-scraping or other unconventionally collected data. *arXiv preprint arXiv:2010.05287*.

Arbia, G., Solano-Hermosilla, G., Micale, F., Nardelli, V., and Genovese, G. (2020). Post-sampling crowdsourced data to allow reliable statistical inference: the case of food price indices in nigeria. *arXiv preprint arXiv:2003.12542*.

Baltagi, B. H., Egger, P., and Pfaffermayr, M. (2007). Estimating models of complex fdi: Are there third-country effects? *Journal of econometrics*, 140(1):260–281.

Bennett, R. J., Haining, R. P., and Griffith, D. A. (1984). The problem of missing data on spatial surfaces. *Annals of the Association of American Geographers*, 74(1):138–156.

Cavallo, A. and Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, 30(2):151–78.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. Academic Press.

Flores-Lagunes, A. and Schnier, K. E. (2012). Estimation of sample selection models with spatial dependence. *Journal of applied econometrics*, 27(2):173–204.

Griffith, D. A., Bennett, R. J., and Haining, R. P. (1989). Statistical analysis of spatial data in the presence of missing observations: a methodological guide and an application to urban census data. *Environment and Planning A*, 21(11):1511–1523.

Haining, R., Griffith, D. A., and Bennett, R. (1984). A statistical approach to the problem of missing spatial data using a first-order markov model. *The Professional Geographer*, 36(3):338–345.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.

Kelejian, H. H. and Prucha, I. R. (2010). Spatial models with spatially lagged dependent variables and incomplete data. *Journal of geographical systems*, 12(3):241–257.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Muris, C. (2020). Efficient gmm estimation with incomplete data. *Review of Economics and Statistics*, 102(3):518–530.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Smith, T. E. (2009). Estimation bias in spatial models with strongly connected weight matrices. *Geographical Analysis*, 41(3):307–332.

Chapter 4

Spatial weight matrix estimation with maximum likelihood LASSO

4.1 Introduction

Spatial econometric models with predetermined spatial weight matrix have been gaining popularity since Anselin (1988). In empirical studies, the choice of the weight matrix remains a problem for the researchers, contiguity based matrices, geographic or economic distance-based matrices are some popular candidates. To avoid the specification being overly arbitrary, some previous studies have explored methods to estimate spatial weight matrices. Bhattacharjee and Jensen-Butler (2013) proposed a method to infer weight matrix in spatial error models under the assumption that the matrix is symmetric. Beenstock and Felsenstein (2012) use moments from the sample covariance matrix to estimate the weight matrix in spatial lag models under restrictions.

Least absolute shrinkage and selection operator(LASSO) based methods for spatial weight matrix estimation have been utilized in recent years. This type of methods assume sparsity of the matrix, but does not have restrictions over the structure of the matrix or sample size T much larger than spatial dimension n, which are often required for previous methods. Ahrens and Bhattacharjee (2015) propose a two-step LASSO estimator for spatial lag models with potentially large n and reasonable T. Lam and Souza (2019) introduced a LASSO based method which allows a combination of predetermined weight matrix as well as an estimated one.

This study proposes a maximum likelihood LASSO (ML-LASSO) estimator for the spatial weight matrix. Compared to previous studies, our method does not necessarily need independent variables since it does not require instruments. We put forward a cyclic coordinate descent algorithm the optimization and to

prove the effectiveness of the estimator with Monte Carlo experiments. An empirical application is demonstrated with US state-level precipitation data from 1895 to 1997.

In section 2, we define the model used in this study, and the estimation method is explained in section 3. Section 4 presents the Monte Carlo experiments verifying the performance of the proposed estimator and section 5 gives an empirical example using the US precipitation data. Section 6 concludes this study.

4.2 Model

We focus on spatial lag model (Anselin, 1988) in this study, which is given by

$$\begin{aligned} y_{ti} =& \sum_{j=1}^{n} w_{ij} y_{tj} + x_{ti}\beta + \varepsilon_{it} \\ \varepsilon_{it} \sim& N(0, \sigma^2) \end{aligned} \tag{4.1}$$

where y_{ti} is a spatial panel data at time t and spatial unit i for $t = 1, \ldots, T$ and $i = 1, \ldots, n$, w_{ij} is the element of the $n \times n$ spatial weight matrix W at the ith row and jth column, x_{ti} is a vector of independent variables $\{x_{1,ti}, x_{2,ti}, \ldots, x_{k,ti}\}$ and β is the corresponding vector of parameters, and ε_{it} is the error term that follows a normal distribution with an variation of σ^2.

The model can be expressed in the following matrix form

$$Y = (I_T \otimes W)Y + X\beta + \varepsilon \tag{4.2}$$

where Y is a $nT \times 1$ vector of y_{ti}s, X a $nT \times K$ matrix of x_{ti}s, I_T a $T \times T$ diagonal matrix, ε a vector of error ε_{it}s, and $\otimes$ denotes the Kronecker product. With this matrix form, the log-likelihood function of (4.1) is

$$\begin{aligned} \log \ell(W, \beta, \sigma^2) =& -\frac{nT}{2} \log(2\pi\sigma^2) + T \log |I_n - W| - \frac{1}{\sigma^2} e'e \\ e =& (I_{nT} - (I_T \otimes W))Y - X\beta \end{aligned} \tag{4.3}$$

Note that for our purpose W is a parameter, instead of being fixed like in normal maximum likelihood estimation for such model.

4.3 ML-LASSO estimator

We propose a coordinate descent type algorithm to estimate the spatial weight matrix W. First we add the LASSO penalty to Equation (4.3), obtaining the penalized log-likelihood function

$$p \log \ell(W, \beta, \sigma^2) = \log \ell(W, \beta, \sigma^2) - \lambda ||W||_1, \tag{4.4}$$

where λ is the penalty level, and $||\quad||_1$ denotes the L1-norm. To optimize the elements of the spatial weight matrix, we partially differentiate (4.4) with respect to w_{ij}

$$\frac{\partial p \log \ell}{\partial w_{ij}} = \frac{TC_{ij}}{|I_N - W|} + \frac{\sum_{t=1}^{T} y_{tj} e_{ti}}{\frac{\sigma^2}{2}} - sign(w_{ij})\lambda, \tag{4.5}$$

where C_{ij} is the cofactor of w_{ij}; e_{ti} is the tith element of e defined in (4.3), and then solve

$$\frac{\partial p \log \ell}{\partial w_{ij}} = 0. \tag{4.6}$$

Since $|I_N - W|$ is always non-zero by definition, Equation (4.6) equals to

$$\begin{aligned}
&Aw_{ij}^2 + Bw_{ij} + C = 0 \\
&A = -\frac{C_{ij} \sum_{t=1}^{T} y_{tj}^2}{\frac{\sigma^2}{2}} \\
&B = \frac{C_{ij} \sum_{t=1}^{T} y_{tj} e_{tij}^* - \sum_{t=1}^{T} y_{tj}^2 \sum_{k=1,k\neq j}^{n} C_{ik} w_{ik}}{\frac{\sigma^2}{2}} - sign(w_{ij})\lambda C_{ij} \\
&C = -TC_{ij} + \frac{\sum_{t=1}^{T} y_{tj} e_{tij}^* \sum_{k=1,k\neq j}^{n} C_{ik} w_{ik}}{\frac{\sigma^2}{2}} - sign(w_{ij})\lambda \sum_{k=1,k\neq j}^{n} C_{ik} w_{ik} \\
&e_{tij}^* = (\ell_N - w_{i,\setminus j})' Y_t - X_{ti}\beta
\end{aligned} \tag{4.7}$$

where $w_{i,\setminus j}$ is the element of ith row of W with jth element fixed to 0, and solving the quadratic equation yields the solution of the single variable optimization problem. The penalty level λ is chosen by minimizing extended Bayesian information criteria(EBIC) proposed by Chen and Chen (2008)

$$EBIC = -2\log \ell + n_w \log(n+T) + 2\phi \log(n(n-1))) \tag{4.8}$$

where ϕ is a predetermined parameter with a value between 0 and 1. β and σ^2 can be estimated when W is given

$$\begin{aligned}
\beta &= (X'X)^{-1} X'(I_{nT} - (I_T \otimes W))Y \\
\sigma^2 &= \frac{e'e}{nT}
\end{aligned} \tag{4.9}$$

We utilize active set strategy (Nocedal and Wright, 2006) to speed up the optimization process, and the estimation algorithm can be conducted as follows:

(i) Standardize dependent variable Y

(ii) Initialize W and estimate β and σ^2

(iii) Select penalty level λ

(iv) Update all w_{ij} by solving corresponding (4.7), then update β and σ^2

(v) Update all non-zero w_{ij}, and then β and σ^2

(vi) Repeat step (iv) until no new zero elements are estimated in W

(vii) Update all w_{ij}, and then β and σ^2

(viii) If W estimated in Step (v) and (vi) have different zero elements, repeat step (iv) to (vi)

(ix) Repeat Step (ii) to (vi) until EBIC is minimized

Step (i) standardizes Y so that Y has a mean of 0 and a variance of 1. In Step (ii), W is initialized as a zero matrix and β and σ^2 are estimated by (4.9). To obtain the optimal penalty level, we use the covariance matrix adaptation evolution strategy (CMA-ES, Hansen and Ostermeier (1996)), because the function $EBIC(\lambda)$ is not smooth and a good global optimizer is needed. In Step (iii) the value of λ is picked by CMA-ES and the check in Step (ix) is also made by the algorithm. The choice of parameter ϕ is quite important when T is not significantly larger than n, we suggest using $\phi = 1$ when there are strong predictors, 0.5 or lower if there is none. In Step (iv), (v) and (vii), we update w_{ij} cyclically. A random updating scheme is also tested, but despite its computational time disadvantage it does not yield better results. Also, we follow Ahrens and Bhattacharjee (2015)'s practice to apply a threshold by defining values under a certain level θ to be 0 when w_{ij} is updated in (vii), to speed up the algorithm and improve the performance. For data sets with no independent variable X, remove terms with X and β in this section.

The algorithm is implemented with R(R Core Team, 2016), Rcpp(Eddelbuettel and François, 2011), RcppArmadillo(Eddelbuettel and Sanderson, 2014), and rCMA(Konen and Hansen, 2015)[1].

4.4 Simulation

This section presents the performance of the proposed estimator for model (4.1) with simulated data sets. Two different underlying spatial weight matrices used in the simulations are specified by

$$W_{s1,ij} = \begin{cases} 1 & when \quad |j-i| = 1 \\ 0 & otherwise \end{cases}, \qquad W_{s2,ij} = \begin{cases} 1 & when \quad j-i = 1 \\ 0 & otherwise \end{cases}, \tag{4.10}$$

then row-normalized by scaling the elements to ensure every row has a sum of ρ, where ρ is a value between -1 and 1 controlling the strenth of the spatial correlation. W_{s1} has $2(n-1)$ non-zero elements on its super-diagonal and sub-diagonal; W_{s2} has $n-1$ non-zero elements on its super-diagonal . W_{s2} is more sparse

[1] We modified the library since it had a bug making it impossible to do one-dimensional optimization

than W_{s1} but we expect that being triangular makes it harder to estimate the direction of the effects.

Table 4.1 reports the performance of the estimator under a series of specifications, which are combinations of $n = \{30, 50, 100\}$, $T = \{50, 100, 150\}$, $W = \{W_{s1}, W_{s2}\}$. The dependent variable Y is simulated from

$$Y = (I_T \otimes \rho W)^{-1}(X\beta + \varepsilon) \tag{4.11}$$

with a vector of independent variable X generated from $n(0, 1)$ and the corresponding β equals 1, error ε from $N(0, 0.1)$, and $\rho = 0.7$. The table reports the rate of true positive, false negative, false positive and true negative in 100 iterations at element level. We can see the performance improves when T increases across all n and specification of weight matrix, in terms of the decrease of false positive and false negative. Similar improvement can be seen when n increases, which is a result of increased sparsity. The estimator works well under both weight matrix specification, although results with W_{s2} have more false positive proportionally despite being more sparse.

Table 4.2 presents the performance of the estimator under the same specifications as Table 4.1, except without a independent variable.The dependent variable Y is simulated from

$$Y = (I_T \otimes \rho W)^{-1}\varepsilon. \tag{4.12}$$

Under W_{s1}, the estimator has reasonable performance but not as good as the results in Table 4.1, with more false negatives and false positives when other parameters stay the same. We still see the increase of n and T bring similar improvements. On the other hand, the asymmetrical nature of specification 2 of the weight matrix brings trouble to the estimator. Figure 4.1 shows the frequency each element being identified as non-zero under certain specification. Figure 4.1d has sub-diagonal and super-diagonal elements estimated as non-zero at a similar frequency, indicates that under W_{s2} the estimator can't identify the direction of the spatial effect in this case. Furthermore, Table 4.2 shows increasing n or T can not solve the problem.

Table 4.3 reports how the strength of the spatial effect impact the performance of the estimator. False-positive and false-negative rate both decrease with the increase of ρ, except when under W_{s1} and no independent variable. As expected, the estimator performs better when the spatial effect is strong.

Table 4.4 presents the performance of the estimator with and without post-LASSO estimation. Ahrens and Bhattacharjee (2015) illustrates that the performance of their two-step LASSO estimator can be enhanced by post-LASSO estimation, reducing the number of false positives. We implement our post-LASSO estimator by maximizing Equation (4.3) with β, σ^2 and non-zero elements in W identified by our ML-LASSO, with R(R Core Team, 2016), Rcpp(Eddelbuettel and François, 2011), RcppArmadillo(Eddelbuettel

and Sanderson, 2014), reach(Schmidt, 2016), and MATLAB Optimization Toolbox(MATLAB Optimization Toolbox, 2020). The W_{s2} with no independent variable case is dropped for this simulation since the LASSO result has too many false negatives which can not be improved by post-LASSO. Table 4.4 shows marginal improvement for specifications with independent variables, and deterioration for cases without them. As a result, we believe post-LASSO estimation is not necessary for our ML-LASSO estimator for identifying non-zero elements.

W spec.	n	T	True Pos. rate	False Neg. rate	False Pos. rate	True Neg. rate
W_{s1}	30	50	0.9997	0.0003	0.0015	0.9985
W_{s1}	30	100	1.0000	0.0000	0.0003	0.9997
W_{s1}	30	150	1.0000	0.0000	0.0002	0.9998
W_{s1}	50	50	0.9999	0.0001	0.0009	0.9991
W_{s1}	50	100	0.9999	0.0001	0.0002	0.9998
W_{s1}	50	150	1.0000	0.0000	0.0000	1.0000
W_{s1}	100	50	0.9997	0.0003	0.0006	0.9994
W_{s1}	100	100	0.9999	0.0001	0.0001	0.9999
W_{s1}	100	150	1.0000	0.0000	0.0000	1.0000
W_{s2}	30	50	1.0000	0.0000	0.0019	0.9981
W_{s2}	30	100	1.0000	0.0000	0.0012	0.9988
W_{s2}	30	150	1.0000	0.0000	0.0007	0.9993
W_{s2}	50	50	1.0000	0.0000	0.0013	0.9987
W_{s2}	50	100	1.0000	0.0000	0.0007	0.9993
W_{s2}	50	150	1.0000	0.0000	0.0004	0.9996
W_{s2}	100	50	1.0000	0.0000	0.0007	0.9993
W_{s2}	100	100	1.0000	0.0000	0.0004	0.9996
W_{s2}	100	150	1.0000	0.0000	0.0003	0.9997

Table 4.1: ML-LASSO simulation results with independent variables, EBIC parameter $\phi = 1$, threshold value $\theta = 0.05$

W spec.	n	T	True Pos. rate	False Neg. rate	False Pos. rate	True Neg. rate
W_{s1}	30	50	0.9867	0.0133	0.0024	0.9976
W_{s1}	30	100	0.9998	0.0002	0.0003	0.9997
W_{s1}	30	150	0.9998	0.0002	0.0001	0.9999
W_{s1}	50	50	0.9730	0.0270	0.0019	0.9981
W_{s1}	50	100	0.9998	0.0002	0.0003	0.9997
W_{s1}	50	150	1.0000	0.0000	0.0000	1.0000
W_{s1}	100	50	0.9895	0.0105	0.0029	0.9971
W_{s1}	100	100	0.9999	0.0001	0.0003	0.9997
W_{s1}	100	150	1.0000	0.0000	0.0000	1.0000
W_{s2}	30	50	0.7655	0.2345	0.0313	0.9687
W_{s2}	30	100	0.9914	0.0086	0.0388	0.9612
W_{s2}	30	150	0.9976	0.0024	0.0371	0.9629
W_{s2}	50	50	0.2969	0.7031	0.0068	0.9932
W_{s2}	50	100	0.9943	0.0057	0.0230	0.9770
W_{s2}	50	150	0.9978	0.0022	0.0218	0.9782
W_{s2}	100	50	0.9584	0.0416	0.0137	0.9863
W_{s2}	100	100	0.9961	0.0039	0.0114	0.9886
W_{s2}	100	150	0.9988	0.0012	0.0110	0.9890

Table 4.2: ML-LASSO simulation results without independent variables, EBIC parameter $\phi = 0.5$, threshold value $\theta = 0.05$

Independent var.	ρ	W spec.	True Pos. rate	False Neg. rate	False Pos. rate	True Neg. rate
Yes	0.3	W_{s1}	0.0098	0.9902	0.0000	1.0000
Yes	0.3	W_{s2}	0.9986	0.0014	0.0012	0.9988
Yes	0.5	W_{s1}	0.9994	0.0006	0.0022	0.9978
Yes	0.5	W_{s2}	1.0000	0.0000	0.0013	0.9987
Yes	0.7	W_{s1}	0.9999	0.0001	0.0009	0.9991
Yes	0.7	W_{s2}	1.0000	0.0000	0.0011	0.9989
No	0.3	W_{s1}	0.0041	0.9959	0.0000	1.0000
No	0.3	W_{s2}	0.0033	0.9967	0.0000	1.0000
No	0.5	W_{s1}	0.3321	0.6679	0.0008	0.9992
No	0.5	W_{s2}	0.2012	0.7988	0.0045	0.9955
No	0.7	W_{s1}	0.9730	0.0270	0.0019	0.9981
No	0.7	W_{s2}	0.2969	0.7031	0.0068	0.9932

Table 4.3: ML-LASSO simulation results with different ρ, n = 50, T = 50, threshold value $\theta = 0.05$, EBIC parameter $\phi = 1$ if there is a independent variable, $\phi = 0.5$ otherwise.

Independent var.	Post LASSO	W spec.	True Pos. rate	False Neg. rate	False Pos. rate	True Neg. rate
Yes	No	W_{s1}	0.9999	0.0001	0.0009	0.9991
Yes	Yes	W_{s1}	0.9999	0.0001	0.0009	0.9991
No	No	W_{s1}	0.9730	0.0270	0.0019	0.9981
No	Yes	W_{s1}	0.9530	0.0470	0.0019	0.9981
Yes	No	W_{s2}	1.0000	0.0000	0.0013	0.9987
Yes	Yes	W_{s2}	1.0000	0.0000	0.0013	0.9987

Table 4.4: post-LASSO results, n = 50, T = 50, $\rho = 0.7$, threshold value $\theta = 0.05$, EBIC parameter $\phi = 1$ if there is a independent variable, $\phi = 0.5$

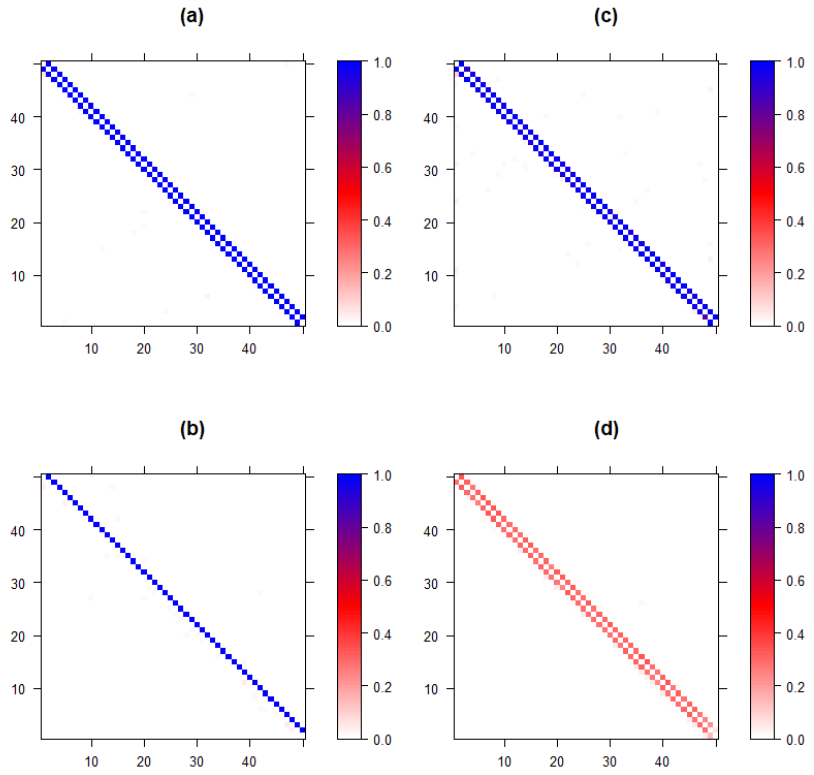

Figure 4.1: Non-zero elements identified by ML-LASSO. (a) W_{s1} with independent variable; (b) W_{s2} with independent variable; (c) W_{s1} without independent variable; (d) W_{s2} without independent variable. n = 50, T = 50, $\rho = 0.7$, threshold value $\theta = 0.05$, EBIC parameter $\phi = 1$ if there is a independent variable, $\phi = 0.5$ otherwise.

4.5 Empirical example

We use the US state-level precipitation data to examine the viability of the proposed estimator in empirical studies. The data set contains monthly weather station data for the continental US from 1895 through 1997, and is obtained from University Corporation for Atmospheric Research(UCAR). To transform the original point data into area data, we divide the stations into groups by a 5×10 grid according to longitude and latitude, see Figure 4.2. We calculate quarterly average precipitation(PPT), elevation(ELE), and latitude(LAT) for each block that has data over the full period. Because operating weather stations changes over time, ELE and LAT are time-dependent.

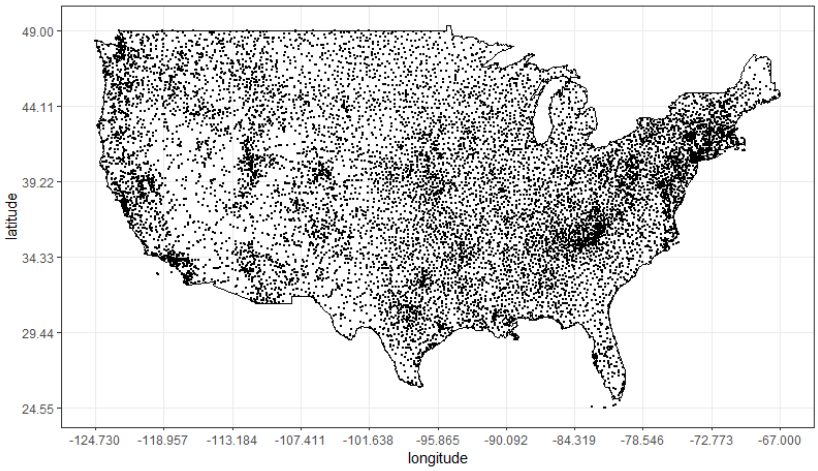

Figure 4.2: Data points, Continental US weather station 1895 - 1997

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
PPT	0.00	9.28	18.66	19.92	28.06	133.20
ELE	6.00	185.60	334.30	583.00	890.90	2085.00
LAT	27.18	34.00	39.09	38.94	42.56	47.10

Table 4.5: Descriptive statistics of US precipitation data

We estimate the spatial weight matrix under two specifications, with and

without independent variables. The model for the first specification is

$$PPT_{ti} = c + \sum_{i=1}^{n} w_{ij} PPT_{tj} + \beta_{elev} ELE_{ti} + \beta_{lat} LAT_{ti} + \varepsilon_{it}, \tag{4.13}$$

and the second one is

$$PPT_{ti} = c + \sum_{i=1}^{n} w_{ij} PPT_{tj} + \varepsilon_{it}. \tag{4.14}$$

Figure 4.3a and Figure 4.3b shows the estimated weight matrix for the two specifications respectively. We choose EBIC parameter $\phi = 1$ for both Model (4.13) and Model (4.14), although the choice of *phi*s do not have a significant impact on the results for this data set. A positive correlation among spatial units and their immediate neighbors can be seen in both figures, which is similar to the first order rook contiguity pattern (see Figure 4.3c).

Table 4.6 compares the goodness of fit among the linear model, spatial lag model (SLM) with a predetermined weight matrix, and post-LASSO result of Model (4.13). We use the post-LASSO result here because we want a comparison among unbiased estimators. The linear model is

$$PPT_{ti} = c + \beta_{elev} ELE_{ti} + \beta_{lat} LAT_{ti} + \varepsilon_{it}, \tag{4.15}$$

and the SLM is

$$PPT_{ti} = c + \rho \sum_{i=1}^{n} w_{SL,ij} PPT_{tj} + \beta_{elev} ELE_{ti} + \beta_{lat} LAT_{ti} + \varepsilon_{it}, \tag{4.16}$$

where ρ is a paramater and $w_{SL,ij}$ are elements of a spatial weight matrix based on first order rook contiguity. The estimation results for these two models are obtained by maximum likelihood. We see the post-LASSO has the best performance among these three, in terms of log-likelihood.

	Linear	SLM	post-LASSO
log-likelihood	-59432.1900	-41190.5700	-35300.4900
Intercept	35.1356	25.9240	5.0754
	(0.6260)	(0.4563)	(0.3234)
β_{elev}	-0.0123	-0.0074	-0.0029
	(0.0002)	(0.0001)	(0.0001)
β_{lat}	-0.2070	-0.3415	0.1264
	(0.0161)	(0.0118)	(0.0083)
ρ		0.6981	
		(0.0062)	
σ^2	116.1140	61.6976	30.9970

Table 4.6: Comparison of estimation results of three models

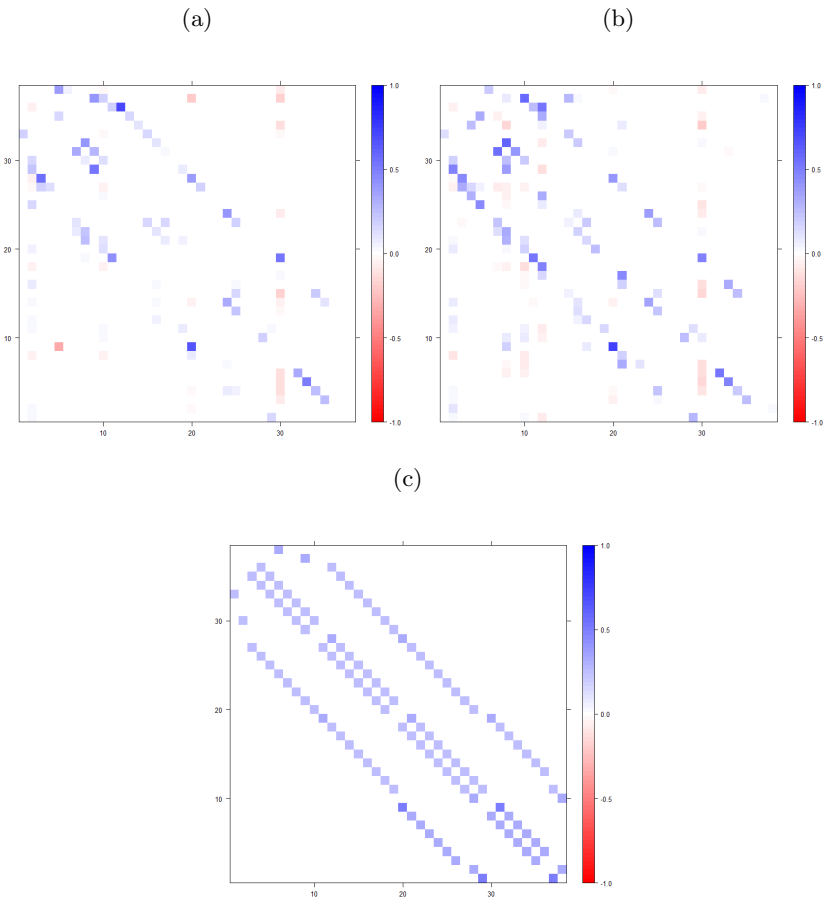

Figure 4.3: (a)Estimated weight matrix, Model (4.13); (b)Estimated weight matrix, Model (4.14); (c)Weight matrix based on first order rook contiguity

4.6 Conclusion

In this study, we introduced an ML-LASSO estimator for spatial weight matrix estimation in spatial lag models. Simulation results show good performance under specifications with arbitrary underlying weight matrices and independent variables, also ones with symmetric weight matrices and no independent variable. Estimation result with US precipitation data indicates the proposed method is capable of identifying spatial patterns in empirical studies.

Compared to previous studies, being able to estimate the weight matrix without independent variables is our advantage. However, it remains a problem for asymmetric cases in the current state, we would like to see further development in the future.

Bibliography

Ahrens, A. and Bhattacharjee, A. (2015). Two-step lasso estimation of the spatial weights matrix. *Econometrics*, 3(1):128–155.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, volume 4. Springer Science & Business Media.

Beenstock, M. and Felsenstein, D. (2012). Nonparametric estimation of the spatial connectivity matrix using spatial panel data. *Geographical Analysis*, 44(4):386–397.

Bhattacharjee, A. and Jensen-Butler, C. (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics*, 43(4):617–634.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063.

Hansen, N. and Ostermeier, A. (1996). Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of IEEE international conference on evolutionary computation*, pages 312–317. IEEE.

Konen, W. and Hansen, N. (2015). *R-to-Java Interface for 'CMA-ES'.*

Lam, C. and Souza, P. C. (2019). Estimation and selection of spatial weight matrix in a spatial lag model. *Journal of Business and Economic Statistics.*

MATLAB Optimization Toolbox (2020). Matlab optimization toolbox. The MathWorks, Natick, MA, USA.

Nocedal, J. and Wright, S. (2006). *Numerical optimization.* Springer Science & Business Media.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schmidt, C. (2016). *reach: Improving interoperability between R and MATLAB*. R package version 0.4.5.